DOCUMENT RESUME

ED 190 609 TM 800 401

AUTHOR Taylor, Hugh: And Others
TITLE Construction and Use of Classroom Tests: A Resource Book for Teachers.
INSTITUTION British Columbia Dept. of Education, Victoria.: Victoria Univ. (British Columbia).
PUB DATE Dec 78
NOTE 50p.: For related document see ED 177 225.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests: Cutting Scores: Decision Making: Educational Objectives: *Educational Testing: Elementary Secondary Education: Item Analysis: Scores: *Teacher Made Tests: *Test Construction: *Test Interpretation: Test Items: Test Reliability

ABSTRACT

This guide is organized into five chapters: (1) an approach to testing--decisions made from test results, types of achievement tests, types and levels of objectives, and test validity, reliability, and practicality: (2) classroom test construction--planning, item banks, item writing, assembly, administration, and scoring: (3) test analysis-item analysis and interpretation of difficulty and discrimination indices: (4) interpretation of test performance--frequency distributions of scores, measures of central tendency, standard deviation, score interpretation, reliability, and measurement error: and (5) procedures for setting standards--minimally acceptable performance, determination of borderline group, and classification errors. An eight-item bibliography and a 55-item glossary are appended. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED190609

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Construction and Use of Classroom Tests

A Resource Book for Teachers

TM 800401

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
P. Northover

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

LEARNING ASSESSMENT BRANCH
THE MINISTRY OF EDUCATION
PROVINCE OF BRITISH COLUMBIA

CONSTRUCTION AND USE OF CLASSROOM TESTS:

A RESOURCE BOOK FOR TEACHERS

Prepared by

Hugh Taylor, University of Victoria
R. Nancy Greer, Learning Assessment Branch
Jerry Mussio, Curriculum Development Branch

Learning Assessment Branch
Ministry of Education
Province of British Columbia

December 1978

3

JUN 2 0 1980

PREFACE

Educators are constantly required to make decisions designed to improve the achievement of students. The vast majority of these decisions are made on a daily basis by the classroom teacher. Consequently, an essential aspect of the teaching-learning process in the classroom must be regular monitoring of student progress. A major responsibility of the classroom teacher is to undertake these monitoring activities, to carry them out judiciously, and to use the information for effective planning to meet the instructional needs of individual students.

These monitoring activities can take a variety of forms; for example, informal observation of classroom behavior, student exercises and projects, or quizzes, tests, and formal examinations. Among these activities, the teacher-made test is one of the most important and most frequently used devices for evaluating students. It is the purpose of this booklet to provide classroom teachers and other educators with assistance in the construction of such tests, and in the use of the results.

It is important to emphasize that this booklet is not intended to foster a situation in which a disproportionate amount of school time is devoted to formal testing activities. Rather, any measurement of student progress must be based on a clear understanding of intents and purposes, which in turn must focus on the needs of students. The emphasis in this booklet is placed upon providing practical suggestions that will assist teachers in designing valid and reliable tests, and interpreting and using test results. An attempt has been made to present the practical, 'how to' aspect of testing and to present only those theoretical issues required to provide a rational basis for the procedures suggested. A glossary of technical terms and a reference list of suggested readings has been provided for those interested in pursuing the theoretical issues in more detail. A detailed index has been included to facilitate the use of this book as an easily accessible source of testing information when and as that information is required.

Appreciation and gratitude are expressed to Dr. Hugh Taylor of the University of Victoria for his substantial contribution to the preparation of this resource book. His role as principal author of early drafts and his continued advice and counsel as the final manuscript took form are gratefully acknowledged.

R. Nancy Greer,
Learning Assessment Branch
Ministry of Education

December, 1978

TABLE OF CONTENTS

Chapter		Page
1	AN APPROACH TO TESTING	7
	1.1 The Teacher as a Decision-Maker	7
	1.2 Types of Tests	7
	(a) Norm-Referenced Tests	8
	(b) Criterion-Referenced Tests	8
	(c) Objective-Referenced Tests	8
	(d) Domain-Referenced Tests	8
	1.3 Classroom Tests Based on Objectives	9
	(a) Types of Objectives	9
	(b) Level of Cognitive Objectives	9
	(c) Taxonomy of Cognitive Objectives	11
	1.4 Characteristics of a Quality Test	11
	(a) Validity	12
	(b) Reliability	12
	(c) Practicality	12
2	THE CONSTRUCTION OF CLASSROOM TESTS	13
	2.1 Planning a Test	13
	(a) The Table of Specifications	13
	(b) Determining Test Length	13
	2.2 The Test Item File	15
	2.3 Writing Test Items	15
	(a) Suggestions for Improving True-False Items	15
	(b) Suggestions for Improving Matching Items	16
	(c) Suggestions for Improving Multiple-Choice Items	16
	(d) Suggestions for Writing Short-Answer Items	17
	(e) Suggestions for Writing Essay Questions	17
	2.4 Assembling a Test	18
	(a) Sequencing of Items in a Test	18
	(b) Arranging the Items on a Page	18
	(c) Test Instructions for Students	18
	(d) General Suggestions on Organizing a Test	19
	2.5 Administering and Scoring a Test	19
	(a) Scoring an Objective Test	19
	(b) Scoring an Essay Test	20
3	THE ANALYSIS OF CLASSROOM TESTS	23
	3.1 Analyzing Test Items	23
	3.2 A Classroom Procedure for Conducting Item Analyses	23
	3.3 Interpreting Difficulty Indices	25
	3.4 Interpreting Discrimination Indices	26
	3.5 Item Analysis by a Teacher	27

TABLE OF CONTENTS continued

Chapter		Page
4	SUMMARIZING AND INTERPRETING TEST PERFORMANCE	29
4.1	Organizing and Describing a Set of Test Scores	30
	(a) Constructing Frequency Distributions	30
	(b) Graphing Frequency Distributions	30
4.2	Measures of Central Tendency	32
	(a) The Mean	32
	(b) The Median	32
	(c) Use of the Mean and Median	33
4.3	A Measure of Score Variability	33
	(a) The Standard Deviation	33
	(b) Uses of the Standard Deviation	34
4.4	Interpreting Test Scores	34
4.5	Reliability	36
	(a) Methods of Calculating a Reliability Coefficient — Kuder-Richardson Technique — Saupe Reliability	36
	(b) Interpreting a Reliability Coefficient	37
4.6	Standard Error of Measurement	38
4.7	Suggestions for Improving Reliability	40
5	PROCEDURES FOR SETTING STANDARDS	41
5.1	Setting Standards on a Classroom Test	41
	(a) Procedure 1: Minimally Acceptable Performance	43
	(b) Procedure 2: Borderline — Group	44
5.2	Errors of Classification	44
	BIBLIOGRAPHY AND SUGGESTIONS FOR FURTHER READING	46
	GLOSSARY OF MEASUREMENT TERMS	47

CHAPTER 1
AN APPROACH TO TESTING

1.1 THE TEACHER AS A DECISION MAKER

Tests help educators make decisions in a number of different areas. Some of the more important ones are described below:

Many decisions are **instructional decisions.** For example, a teacher must decide whether the majority of students in the class are sufficiently competent in using a simple mathematical operation or whether a review is needed prior to beginning advanced work. Some instructional decisions relate to individuals when, for example, a teacher must decide what reading level will be most appropriate for Maria in recommending a novel for her enrichment reading.

Other decisions deal with **curricular decisions.** A school or a district might consider increasing the emphasis given to physical activities, athletics and cultural pursuits and may wish to determine the effects of this on the traditional academic areas of the curriculum. Knowing the overall achievement levels before and after changes to the curriculum are made can help administrators judge the effects on achievement in the school district.

Another type of decision educators may be called upon to make are **selection decisions.** A college or university may decide, due to limited personnel or financial resources, to restrict the enrolment in one or more of its popular programs. A uniform testing program along with other relevant data can supply important information that will aid the selection staff in making decisions that will admit the potentially most successful applicants.

Some decisions made by school personnel may be called **classification** or **placement decisions.** These decisions relate to assigning letter grades to students, placing individual students in different grade levels, placing students in different sections within a grade or special classes so they may obtain maximum long term benefit from among the various programs organized within a school or district. Should Fred who has a minor specific learning problem be recommended for small-group instruction part of the time? Should Helen be placed so that special assistance can be given in helping her overcome a speech or language difficulty? School personnel may have to decide whether Bert, whose ability and adaptive behavior appear to be extremely limited, should be placed in a program organized for severely retarded students. School tests and information from other professional personnel often help educators make these important placement decisions.

Finally, tests may aid individuals in making **personal decisions.** Should John plan to go to college or attend some other type of post-secondary institution? Given Joan's particular measured interests, abilities and temperament, should she plan on becoming a primary teacher? These kinds of questions can be answered more realistically if individuals have available personal test data to aid them in their decision making.

1.2 TYPES OF TESTS

A test may be thought of as a procedure for sampling the behavior of an individual or groups of individuals. A single test can measure only a small fraction of a person's knowledge and intellectual skills or abilities and, accordingly, a wide range of tests can be developed. There are many different ways of classifying these tests. Some tests must be administered to one individual at a time, such as the Stanford-Binet Intelligence Test, while others may be administered simultaneously to a large number of students such as the Canadian Tests of Basic Skills. Other tests, such as those developed for the British Columbia Learning Assessment Program, are often administered to only a sample of students across the province; the use of statistical theory enables researchers to determine how well all students

would have performed had everyone taken the tests. Tests developed by classroom teachers and used within their particular classrooms are called teacher-made or informal tests, while tests that have been developed by testing companies and which may be administered in a uniform manner in classrooms across the nation are called standardized or published tests. Other classifications such as readiness tests, mastery tests, diagnostic tests and others will be found in the Glossary at the end of this manual.

Recently, it has been popular to think of tests in terms of four major classifications: **norm-referenced tests, criterion-referenced tests, objective-referenced tests** and **domain-referenced tests.** The major difference between the four types relates to how the test results are interpreted. Other differences include the range of difficulty of the test items and the extent of the behaviour sampled by the test.

(a) Norm-Referenced Tests refer to the typical standardized test and many classroom tests where a student's score is judged in terms of how it stands in relation to the scores of other students who wrote the test (the norm group). A common method of giving meaning to an individual's raw score is to convert it into a percentile rank which defines the percentage of pupils who obtained a score less than or equal to the student's score. For example, if Fred's score is such that 75% of the students in the class (or norm group) obtained a raw score less than or equal to Fred's score, we would say that Fred's percentile rank was 75. This value gives his relative performance in terms of his rank in a standard norm group of 100 students.

(b) Criterion-Referenced Test is a label that is commonly used to refer to tests whose scores are interpreted primarily in terms of a pre-determined standard (usually percent correct), in contrast to comparing the scores to norms or to class performance, as with norm-referenced tests. For example, the minimum score for passing the written exam required to obtain a driver's license might be set at 90%. Certain criterion-referenced tests sold by test publishing companies require the student to obtain a score of 80% correct before starting a new unit of subject matter. The questions appearing on such a test are selected to be representative of a clearly defined domain of learning outcomes, and in this way the score is taken to be representative of the student's present status with respect to those outcomes. Criterion-referenced tests are usually shorter and cover a much more limited amount of content than norm-referenced tests. They are most useful when the determination of a student's level of mastery is the main purpose of the testing activity.

(c) Objective-Referenced Tests are very similar to criterion-referenced tests in that the questions appearing on both are selected because they relate to rather narrow, highly specific learning objectives. Both contain items that measure clearly defined objectives, but objective-referenced tests differ from criterion-referenced in that they have no pre-determined performance standard associated with the scores. Their purpose is to survey the tasks that students can perform in different areas of the curriculum. Administered periodically, these tests, or the individual test items, provide useful information for assessing the curriculum and for determining general educational progress. Examples of objective-referenced tests may be found in the various reports issued by the Learning Assessment Branch of the British Columbia Ministry of Education.

(d) Domain-Referenced Tests are used to estimate performance on a universe of items similar to those used on the test. As such, the content area of the test is rather explicitly defined such as, for example, word recognition ability at the primary level or reading comprehension ability at the intermediate level. A large pool of items is developed for the domain and items are randomly sampled from the pool for placement on a particular test. Scores are reported as the percentage of items that a student could get correct in the total pool.

It should be noted that individual items on the four types of tests can be quite similar both in structure and in content. As such, it is sometimes difficult to differentiate between the tests on appearance alone. The major differences are related to how the scores are interpreted as well as to the usefulness of the test results in making various types of decisions. For example, a norm-referenced test is designed primarily to allow for comparisons to be made between individuals and groups of students. Items appearing on these tests have been selected because they have been found to maximize small

differences between students. Any questions that are ineffective at detecting small differences between the achievement levels of students are eliminated during the test development phase.

Objective-referenced and criterion-referenced tests, on the other hand, are primarily concerned with content coverage. Items are selected or rejected on the basis of whether or not they are judged to measure a component of the knowledge or skills specified in the learning objectives to which the tests are referenced. These tests provide valuable information concerning what a student can and cannot do. However, they tend to be less efficient than norm-referenced tests when the scores are to be used to detect small differences between the students for comparative purposes. It is extremely important, then, to ensure that the purposes of testing are clear before a test is developed or selected.

1.3 CLASSROOM TESTS BASED ON OBJECTIVES

Teachers use a wide variety of procedures other than paper-pencil tests to assess the students' progress. These include checklists for judging the students' performance on certain physical activities in the gym, judging techniques in playing the clarinet, evaluating the product in a Home Economics laboratory or woodwork shop, and judging an oral report in a Social Studies class. However, most teachers develop their own paper and pencil tests to survey the students' knowledge and intellectual skills. These teacher-made tests form one of the most important techniques for evaluating students' progress in schools today. It is therefore important that teachers base the construction of their own tests on principles recognized as basic among educational measurement specialists.

To a large extent, formal education is a rational process. Teachers first plan learning goals (objectives) for their students. Next, they attempt to arrange conditions in the classroom that will help the students reach the stated goals. Lastly, they evaluate both the students' progress and learning conditions for the purpose of making adjustments in the curriculum or planning more effective learning conditions in the future.

(a) Types of Objectives: Testing procedures should be based on appropriate learning objectives for students. It is convenient to think of student objectives in terms of three major types: cognitive (thinking), psychomotor (physical activities) and affective (emotional) development. Of course, human behaviour usually cannot be neatly classified into just one of the three categories. For example, think of a gymnast performing on the uneven bars. Obviously a great deal of large and small muscular development (psychomotor learning) has taken place prior to the performance of the athletic event. However, one can also imagine that a tremendous amount of concentrated thought (cognitive learning) was needed to perform the very complicated movements. Also, it is obvious, especially when the performance is completed by a successful dismount, that the whole activity is accompanied by tremendous pleasure (affective learning). Thus, for a comprehensive evaluation of student learning, all three types of objectives should be considered.

(b) Level of Cognitive Objectives: Most school testing deals with three levels of objectives in the cognitive domain. The first level of objectives, called **long term,** gives direction to the educational enterprise. These are goals that laymen usually recognize as important as well as being the proper responsibility of the school. Examples include the following:

1. to acquire the skills of reading
2. to acquire the skills of writing
3. to understand our number system and its use in practical situations
4. to acquire knowledge of science as related to everyday life
5. to develop skills and knowledge for healthful living

Important as these long term goals are for curriculum work and for guiding the overall progress of education, they are of limited use for teachers in their daily planning and testing activities. Other objectives that operationalize the long term goals are the most helpful types for teachers. These include second level objectives called **general instructional objectives** and the third level objectives which are called **specific learning outcomes.** The general objectives refer to the major objectives that describe the intellectual activities and subject matter content which the teacher is trying to promote. Most

norm-referenced tests are based on from three to five general instructional objectives. The specific learning outcomes, subsumed under the general instructional objectives, express student behavior in specific terms. Table 1.1 contains some examples of general instructional objectives and their accompanying specific learning outcome: Note that both types of objectives begin with a verb and that the verbs associated with the specific outcomes are rather easy to interpret, particularly in terms of how students are to respond after learning has taken place.

Table 1.1

Examples of General Instructional Objectives (G.I.O.) and Specific Learning Outcomes (S.L.O.)

Grade 5 Science

G.I.O.: Apply the concept of sound reflection and echoing to predict the soundproofing quality of certain materials.

S.L.O.:
1. Describe how sound travels from its source to the ear.
2. Illustrate, using a diagram, what happens when we hear an echo.
3. Explain the difference between porous and non-porous materials with reference to sound waves.
4. Categorize a list of materials into those which would or would not be suitable for sound proofing.
5. List three possible situations in which sound proofing might be used.

Grade 9 Foods and Nutrition

G.I.O.: Understand the chief role of food nutrients in the body, using Canada's Food Guide as a reference.

S.L.O.:
1. Identify nutrients in Canada's Food Guide.
2. List the valuable sources of each nutrient.
3. Given a list of nutrients found in Canada's Food Guide, select the corresponding deficiency diseases if nutrients are lacking.
4. Plan a balanced meal using Canada's Food Guide.
5. Explain the relationship between physical development and activity during adolescence and the need for adolescents to make proper food choices.

Grade 7 Social Studies

G.I.O.: Understand some specific facts and concepts about Egyptian cultural history.

S.L.O.:
1. Summarize the important influence the Nile had on Egypt's development.
2. Name or describe the pharaoh's tomb contents.
3. State the purpose or purposes of the pyramids.
4. Describe an Egyptian home in the time of the pharaohs.
5. Explain the general beliefs of the early Egyptian religion.

Grade 10 Mathematics

G.I.O.: Understand the facts and principles of multiplying and factoring binomial expressions.

S.L.O.:
1. Use the distributive axiom and the rules of exponents to multiply a given binomial by a monomial.
2. Factor a given binomial that is the product of a monomial and a binomial.
3. Write the product of the sum and difference of two given numbers as a polynomial.
4. Determine whether or not a given binomial is the difference of two squares.

(c) Taxonomy of Cognitive Objectives: During the last few years, considerable effort has been directed towards developing principles for writing and organizing educational objectives. One of the most comprehensive schemes is that proposed by Bloom and his associates in their **Handbook of Formative and Summative Evaluation of Student Learning** (1971). Covering procedures for evaluating learning from pre-school to university, this text should be available as a reference source for all teachers. However, rather than using the rather complicated Bloom procedure, one can adopt a relatively simple and practical way of writing and organizing cognitive objectives. This consists of classifying the general instructional objectives into three main categories, as defined in Table 1.2, along with examples of some verbs appropriate for use with the general instructional objectives and specific learning outcomes.

Table 1.2

Categories of Intellectual Behaviour, Definitions and Sample Verbs

Categories of Behavior	Definition	Sample Verbs for	
		General Instructional Objectives	Specific Learning Outcomes
1. Knowledge	Remember facts, ideas, terms, conventions, methodology, principles and generalizations	Knows	Defines, describes, lists, names, outlines, selects, states
2. Simple Understanding	Interprets, translates, summarizes or paraphrases given material	Understands, interprets, translates	Computes, converts, illustrates, interprets, predicts, rearranges, paraphrases
3. Complex Problem Solving	Solves problems by transferring prior knowledge and/or learned behaviour to new situations, analyzes complex situations, creates unique products, makes judgements based on established standards or sets new standards	Applies, analyzes, writes, judges	Appraises, composes, creates, criticizes, discovers, infers, relates, solves

The purpose in printing this table is not to suggest that it should be used rigidly but to present a relatively simple model with the hope that it will encourage teachers to consider carefully their own objectives and, in particular, to attempt to organize them into some type of hierarchical arrangement. By using this model, a teacher will often discover that too much emphasis is being placed on, say, remembering facts, and too little emphasis on the more complex skills of analyzing and judging. Well organized objectives can simplify the evaluation process and also aid in the overall planning of a test through the use of a table of specifications which will be discussed in detail in a later section.

1.4 CHARACTERISTICS OF A QUALITY TEST

One must make sure that the information provided by test scores actually serve the purposes for which the test was designed. Decisions made on the basis of tests will be valuable only to the extent that the inferences made from the test scores are appropriate. The process of judging what may properly be inferred from an achievement test score is known as determining the **validity** of the test.

(a) Validity: Tests can be valid for one type of decision, but invalid for another. When speaking of test validity, the question must be 'Valid for what?'

There are several types of validity. Probably the most important is **content validity** for the types of testing most frequently carried out in the classroom. Content validity can be demonstrated by showing that the behaviours performed in testing constitute a representative sample of behaviours specified by the objectives of the unit or course. In other words, to be valid, a unit test in Social Studies must measure the kinds of skills which are taught in the unit. Each item on the test should measure one or more course objectives and all items taken together should represent an appropriate sample of the total unit or course objectives. Procedures for studying the content validity of the test will be discussed later under the topic of Table of Specifications.

Construct validity refers to the degree to which the test actually measures what it proports to measure. While this type of validity is often more critical when selecting a published test than when constructing a classroom test, its importance should not be overlooked entirely. For example, if the Social Studies test referred to above required a reading level far in excess of many of the children answering it, the test would have low construct validity as it would be measuring differences in reading ability rather than differences in knowledge of Social Studies. Similarly, a Mathematics test given under conditions of extreme time pressure and tension would probably be reflecting differences in test anxiety rather than differences in mathematical ability.

Criterion-related or predictive validity is paramount when the test scores are being used to assess the student's likelihood of success in some future undertaking. For example, a test that is used to select students for a special program for gifted and talented children must have high predictive validity. That is, it must be able to identify those children most likely to be successful and well-suited to such a program. With classroom tests, criterion-related validity becomes important if the test scores are being used to indicate a student's readiness for a subsequent unit of instruction.

(b) Reliability, another characteristic of a quality test, refers to the degree to which score differences within a class are attributable to true differences rather than chance differences in student achievement. If Maria scored 80 on a Mathematics test and Fred 70, can we really be sure that Maria is a better student? If a test has high reliability, we can be confident that Maria has, in fact, done better in the course or unit than Fred has. If a test has low reliability then score differences should be ignored. Many helpful suggestions for increasing the reliability of test data are discussed in Chapter 4 of this booklet.

(c) A third characteristic of a quality test is **Practicality.** To be practical, a test, in its construction, administration, scoring and interpretation must make the most effective and efficient use of both student and teacher time. Suggestions for improving these aspects of practicality will be discussed in the following chapter.

CHAPTER 2
THE CONSTRUCTION OF CLASSROOM TESTS

In this chapter a number of procedures to be applied when developing a classroom test or examination are described. On first reading these procedures may appear to be needlessly detailed and time-consuming. However, it is attention to these very details during the developmental stages which results in tests that are reliable, valid, practical, and most importantly from the point of view of the student, "fair" evaluations of performance. The more important the decision which is to be made using the test results, the more critical it is that these procedures be applied.

2.1 PLANNING A TEST

(a) **The Table of Specifications:** A test should be planned well in advance of the time it is to be administered. Adequate planning for the test, while time consuming, will yield considerable savings in time when the test is to be marked and the results interpreted. It is helpful for initial planning to have a detailed outline of the content of the course and a list of the various objectives which are to be tested. It is important, then, to develop what is called a **table of specifications** for the test. A table of specifications is a two-way chart showing the **content categories** of the proposed test along the vertical axis with three **intellectual levels** placed along the horizontal axis. The outer section of the table shows the percentage of test items that are related to any particular row or column heading while the inner cells (once the test is printed) list the actual item numbers on the test. An example of a table of specifications for a unit test in mathematics, composed of 40 items, is shown in Table 2.1.

Table 2.1
Table of Specifications
Unit One Problem Solving Mathematics 9

Content	Cognitive Level			% of Total
	Knowledge	Understanding	Problem Solving	
Algebraic Expressions and Equations	1, 5, 6, 7	3, 4, 8, 9, 12	2, 10, 11	30%
Word Problems	13	14, 15, 16	17, 18	15%
Problems of Two Related Unknowns	20, 21, 22, 23	26, 28, 29	19, 24, 25, 27, 30	30%
Using Formulas		34, 36, 37	32, 32, 33, 35, 38, 39, 40	25%
% of Total	28%	20%	46%	100%

Numbers in the cells refer to the numbers of the test items on the test corresponding to the particular cell.

The percentage of test items within any row or column in the table of specifications is determined by the teacher, based upon such considerations as the amount of time devoted to the different content areas and the emphasis placed on the different types of intellectual behavior that were stressed during the course.

(b) **Determining Test Length:** Once a table of specifications for a test is drawn up, it becomes necessary to decide how many test questions to include for each cell of the table. What constitutes a sufficient number of test questions centers around two issues.

The first of these concerns content coverage. If you will be giving the test to determine whether the student has mastered the content as stated in the specifications, it is important that the questions constitute a fully representative sample of those behaviors. For example, if the objective is to demonstrate ability to add single digit numbers, what will constitute a sufficient number of questions will be influenced by issues such as: how many numbers are to be added; whether negative values are to be included; whether numbers are to be placed both horizontally and vertically on the page; whether multiple choice as well as student supplied response formats are to be included, and so on. Whether a student is judged to have mastered this objective clearly will be influenced by the types of questions that are used on the test.

Figure 2.1

Test Item Card with Item-Analysis Data Entered

FRONT

No.____ Grade____ Course Math 9 Curr. Guide 19____ pp.____

Content or Topic Simple Interest Formula Cognitive Class 4

Objective Solve problems concerning simple interest.

Reference _____

31. A store owner borrowed a sum of money for 9 months. He paid back $1800 which included the amount he borrowed in addition to his interest at 12%. How much did he borrow?

A $1651
B $1638
C $1584
D $1605

(A) B C D E

BACK

Date	Oct./78
Item No.	31
N	30

OPTIONS	Upper	Lower	Upper	Lower	Upper	Lower
A	0	2				
B	0	2				
C	6	3				
D	2	1				
E						
Omit	2	2				
Diff.	.10					
Disc.	-					

Oct./78
Very difficult, review
supply formula with
stem?

14

The second issue central to determining test length is the degree of confidence that is to be placed in the scores achieved on the test. Suppose that you had given the 40 item test of Problem Solving referred to in Table 2.1 to a Math. 9 class. You want to use this test to determine whether students had mastered this unit of the course. You decided that a score of 32 out of 40 would be accepted by you as mastery level performance. The question now is "How likely am I to make errors and misclassify masters as non-masters or non-masters as masters using the scores on this test?" Test scores are fallible. Students may answer questions correctly by guessing when they have not mastered the content. Conversely, those who do know the content may make mistakes through inattention, fatigue, or other causes unrelated to their true level of ability. The important point to be made here is that the length of the test will have an influence on the confidence that can be placed in the scores. In general, as the number of questions the student is required to respond to for a given objective increases, the risk of drawing an incorrect conclusion about the student's level of mastery of that objective decreases.

2.2 THE TEST ITEM FILE

Once the table of specifications has been designed, the teacher has either to compose the test items or select them from an item file. A convenient way for developing a test file is to write each item on a 8½" x 5" test item card, similar to the one shown in Figure 2.1. Space is reserved on the back for notes and recording item analysis data.

Developing a test item file is a difficult task which can be less arduous if several teachers who teach the same subject in a secondary school or who teach the same grade level in an elementary school can work together cooperatively. Each can contribute items and make use of those provided by other members of the group. Also each can provide editorial comments and suggestions for item revision which can greatly improve the validity of the items.

2.3 WRITING TEST ITEMS

Items for teacher-made tests are usually classified into two major types, depending upon whether the student selects the answer from a number of options or whether the student actually supplies the answer. Examples of **selection-type items** include True-False, Matching and Multiple-choice while **supply-type items** include the Short Answer and Essay.

The following suggestions for writing various types of test items should provide a greater assurance that the items will actually test what is intended by the teacher, thus increasing the validity of the test.

(a) **Suggestions for Improving True-False Items:**

1. Avoid trivia; develop items which require students to think with what they have learned rather than simply to recall it.
2. Simplify statements as much as possible; avoid double negatives or unnecessarily involved sentences.
3. Make each item deal with a single definite idea. The use of several ideas in each statement tends to be confusing and the item is more likely to measure reading ability rather than achievement.
4. Avoid making true statements longer than false statements.
5. Have an approximately equal (but not exactly equal) number of true and false statements and vary the proportions from test to test.
6. Randomly arrange true and false items; check to be sure there is no inadvertent pattern.
7. Be sure the items can be unequivocally classified as true or false.
8. Avoid the use of statements extracted from text books. Out of context such statements are often ambiguous.
9. Beware of such specific determiners as **all** or **none** (clues to a false statement) or **generally, usually, some, sometimes** (clues to a true statement).
10. Make the method of response as simple as possible, such as circling a capital T if the statement is true or capital F if false.
11. Give careful consideration to whether another type of question format can be used (i.e. multiple choice). Unless extremely well-written, true-false questions can produce results of low reliability.

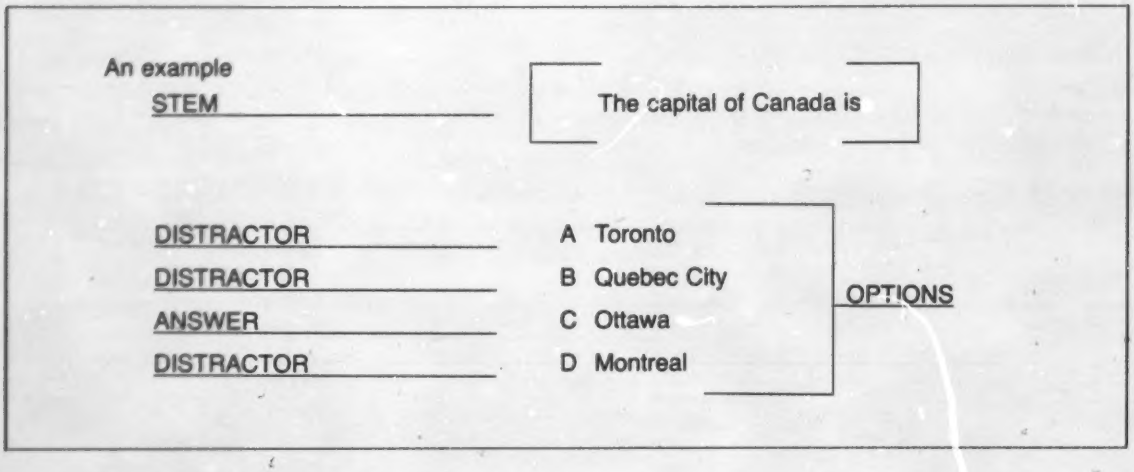
(b) **Suggestions for Improving Matching Items:**

1. The longer, more complex statement should be used as premises and placed in column on the left, the shorter statements or responses on the right. Each item in the left column should have a test number; responses should be preceded by letters. Each column should be given a title.
2. Directions should specify the basis for matching and should indicate whether responses should be used once, more than once or not at all. Use illustrations whenever possible.
3. The premise and response columns should constitute homogeneous lists, each grouped around a single concept; for example,

events and causes	events and people
events and dates	terms and definitions
events and places	rules and examples

4. The list of responses should be at least three longer than the list of premises to preclude guessing by elimination (unless directions indicate that each response may be used more than once).
5. The items should include at least five but no more than twelve premises.
6. Each item should be kept on a single line.
7. Responses should be arranged in some order to simplify matching (alphabetically, chronologically, logically).

(c) **Suggestions for Improving Multiple-Choice Items:** A multiple-choice item is composed of two parts: a **stem** that poses a problem and a number of **options** or possible solutions to the problem. Options include the correct answer plus a number of distractors that should appeal to students who are in doubt about the correct answer.



1. The stem should pose a significant, single problem expressed clearly, accurately and completely. The problem should be practical and realistic.
2. State the stem in positive form whenever possible. When negative wording is used, emphasize it by underlining or capitalizing.
3. The stem should be either a direct question or an incomplete sentence. Beginning item-writers tend to produce fewer technically weak items when they use direct questions.
4. As much of the item as possible should be included in the stem. All the information should be relevant to the solution of the problem unless a specific purpose is to measure ability to sort out relevant material.
5. Make options as brief as possible. Instead of repeating words in each option, include them in the stem.
6. The options should be homogeneous. The more homogeneous the alternatives, the more difficult the item will be in that the item tends to measure higher levels of understanding.

7. Options should be of relatively uniform length. Beginning test constructors often include the largest number of words in the correct answer; because of this make sure, when options do vary in length, that the correct answers are not consistently longer than the alternatives.
8. Options should be grammatically consistent with the stem and as nearly parallel in form as possible.
9. Distractors should represent common errors which actually occur in students' thinking. Excellent distractors can be obtained from incorrect responses on short answer, completion, or essay tests. Distractors can serve as important a function in the question as the correct answer in that they can become a starting point for diagnosis of individual difficulties.
10. The correct answer should be the one that competent critics agree is the best. Avoid options that overlap or include each other.
11. Try to avoid using the options "None of the above" or "All of the above". "None of the above" is appropriate as a last option when there is a correct answer as distinguished from a best answer as, for example, in items involving mathematical problems where the answer is a precise quantity. "All of the above" as a last option should seldom be used. If the student recognizes two correct options he can quickly conclude that "All of the above" is the correct answer. Conversely, if he knows that one option is incorrect, then "All of the above" cannot be the answer.
12. Examine your response options to ensure that there is one correct or clearly best answer.

(d) **Suggestions for Writing Short-Answer Items:** Short-answer items include both the direct question and incomplete sentence type. They are useful in the early elementary school grades. Other types such as multiple-choice and essay are more desirable for use with older students.

1. Construct the question so that only one word, phrase, number, or symbol will satisfy the question. Items should permit only one or a few correct possibilities and the scoring key should list all that exist.
2. Use primarily to measure factual knowledge or to cover large amounts of material over a brief testing time.
3. Avoid the use of verbatim material from a text. The use of the exact words of a text encourages rote memory or parrot-like learning.
4. Avoid mutilated statement. Too many blanks make the question meaningless or impossible to answer. Blank out only key words or phrases.
5. Construct all blanks of a standard length to avoid giving the student clues to the correct answer, and allow enough space to permit a legible answer. Arrange the spaces, usually on the right side of the page opposite the close of the sentence, for convenience in scoring.
6. Specify the degree of precision expected in computational problems.
7. Allow one point for each blank correctly filled.
8. Avoid grammatical clues to the right answer. Write the indefinite article as a(n).

(e) **Suggestions for Writing Essay Questions:**

1. Restrict the use of essay items to the function for which they are uniquely suited. The essay item appears to be of particular value in courses such as English composition and journalism where developing the student's ability to express himself in writing is a major objective. It is also well suited to advanced courses in any subjects where critical evaluation and the ability to assimilate and organize large amounts of material constitute important instructional objectives.
2. Phrase the essay item in a manner that calls upon the appropriate content and intellectual levels within the cells of the table of specifications. Ask yourself continually, "Does this item bring out the information that I want?"
3. State essay questions so that they present a clear, definite task to the student. Since essay questions are to be primarily used to measure understanding and complex problem solving, they are more apt to do so if they start with terms such as "Why", "Describe", "Explain", "Compare", "Interpret", "Analyze", and "Criticize".
4. From a student's point of view, essay tests are very time consuming. Make sure the test does not include so many items that the student does not have sufficient time to consider each one carefully before answering or to review his responses and make any necessary revisions. It is helpful to indicate after each test item how much time should be spent on it as well as how much it contributes to the total score on the test.

5. In general, you should not offer a choice of essay items, particularly in a content field such as Science. For example, if you present six items and ask the student to choose three, you will not have a common basis upon which to evaluate different individuals within the class. If the test data are to be used for grading purposes, they must be based on the same task or set of test tasks. The use of optional questions is, in reality, the administration of several tests of unequal difficulty. However, the practice of allowing students a choice is justified if the purpose of testing is to measure writing effectiveness rather than subject matter acquisition. It allows students to select questions best suited to their writing skills and to avoid the possible frustration of having to write on an unfamiliar topic.
6. When constructing the essay test, it is helpful to define the direction and scope of the response. This can be done by asking a "topic question", then adding several subsidiary questions of a more specific nature. Structuring the question allows all students to attack the same problem and also aids the teacher in preparing answers to be used as keys in scoring.
7. Students require numerous experiences in expressing their ideas in writing. The use of unstructured questions is justified for this task provided the responses are read critically and returned to the students for the purpose of formative evaluation.

2.4 ASSEMBLING A TEST

After items have been written or chosen to assess the various cells in the table of specifications, decisions must then be made concerning the best way to arrange them within the test booklet. The following suggestions should prove helpful for this purpose.

(a) **Sequencing of Items in a Test:**

1. All items of the same format (style) should be grouped together. As each format requires a different set of directions, grouping items makes it possible to have a clear set of directions that will apply throughout that section of the test. It also contributes to effective test taking, since the student maintains a uniform mental set or approach throughout the section. Finally, it tends to simplify the scoring and the analysis of the results.
2. Within each section, it is appropriate to group items according to the sequence in which the material was presented. This makes the student more comfortable as he proceeds through the test. It also facilitates discussion of the test after it has been marked and returned to the student. An alternative arrangement would be to begin each section with very easy items and then progress in difficulty, the purpose being hopefully to instill confidence in the student early in the testing period. There is nothing more discouraging for a student than to begin a test and find he cannot answer the first group of questions. A possible shortcoming of this arrangement of test items from easy to difficult, however, may be the lack of any logical sequence of ideas as the student progresses through the test.

(b) **Arranging the Items on a Page:**

1. A complex set of items that use a common diagram or a common set of responses should be arranged within the booklet in such a way as to avoid the necessity of flipping pages back and forth. Diagrams should be placed above the items in order to avoid a break in the student's reading continuity between the stem and the options.
2. Multiple-choice items often are arranged in two vertical columns on the page. This double-column format makes the test easier and faster to read and also saves space as more items are included on a page. Within a multiple-choice item the options should be placed in a column rather than in paragraph sequence.
3. A Multiple-choice item should be printed so that there is no split in the middle of the question or option at the end of a column on a page.
4. Arrange the items on a page to ensure easy reading and analysis by the students. Different sections of a test should be set off by extra spaces or a line.

(c) **Test Instructions for Students:**

1. The cover page should list the number of items on the test and the total number of pages.
2. Provision should be made on the test or answer sheet for the student's name, class or subject, date, teacher.

3. Indicate the value of each item and the suggested time that should be allowed to answer the question.
4. Each item format should have a specific set of directions; e.g. in the multiple-choice section, the students should be advised as to whether or not they should guess at the correct answer. As a general rule it is not worth the time and effort to use a correction-for-guessing formula with classroom tests.
5. For selection-type tests (i.e. multiple-choice, matching) at the elementary school level, students should be given examples and/or practice exercises so that they can see exactly how to handle the format of the questions.

(d) **General Suggestions on Organizing a Test:**
1. The test should be carefully proofread before it is reproduced. It is often very worthwhile to have it reviewed independently by a colleague who teaches the same subject.
2. Errors found on a test after it has been printed should be pointed out to students before they begin to answer the questions.
3. Items should be numbered sequentially throughout the entire length of the test.
4. The test paper should be of good quality and the legibility of the test must be satisfactory from the viewpoint of the type size, adequacy of spacing and clarity of printing.
5. Students should be given advance notice of all important tests. They should be aware of the amount of material to be covered and have some indication of the length of test that can be expected. They should also be aware of the importance the teacher is placing on a test and of the influence the score will have on the final grade.

2.5 ADMINISTERING AND SCORING A TEST

1. Young students should be given practice in taking tests, particularly if unfamiliar types of items or ways of asking questions are used, e.g., analogy items.
2. It is important that the testing room should be as conducive as possible to concentrations to the task at hand. Ventilation and lighting should be checked for adequacy.
3. Tests should be administered in a way which will allow all, or nearly all, students sufficient time to finish. Students should know what they are permitted to do if they finish early.
4. If possible, there is value in announcing the entire term's schedule of tests in advance, particularly at the secondary school level.
5. Careful proctoring by the teacher is the single most effective method to minimize cheating by the minority of students who might be tempted to resort to it. Depending upon the usual arrangement of desks or tables, it may be desirable to re-arrange these during tests to minimize the temptation to cheat.
6. A well-designed test should result in few questions from students during its administration. Deal with any queries from individual students quietly and quickly. Avoid addressing the class as a whole once the test has begun.

(a) **Scoring an Objective Test:** The following suggestions apply to objective tests made up of items for which the correct answer is set in advance of testing so that scores are unaffected by the opinion or judgement of the scorer.
1. In classroom tests where students are given sufficient time to answer every item on the test, the total score is the number of items the student has answered correctly. Giving students negative scores for certain types of errors, half scores or double scores for relatively unimportant or important items is inconvenient, time consuming and largely wasted effort. If a Table of Specifications was used to ensure that the number of items for various intellectual behaviors and content are in the desired proportions, then the appropriate weighting has already taken place as the test was composed.
2. When the number of students writing the test is 30 or less, it is appropriate to have the students answer the questions directly on the test paper by providing spaces for recording the answers. An unused test booklet with the answers included can then serve as a key. To simplify the mechanics of marking, strip keys can be prepared easily by cutting the columns of answers from the master copy of the test and mounting them on strips of cardboard cut from manilla folders.

3. When the number of students to be tested is large or if detailed analysis of the test results is to be made, it is recommended that separate answer sheets be used. A blank sheet with the holes punched out where the correct answer should appear can be laid over each student's sheet. A convenient way of developing a key is to cut the answer sheet into narrow strips, one answer column to a strip; punch out, then reassemble the entire answer sheet by taping on the back side, being careful not to cover up any of the punched out holes. As each paper is scored, it is useful to mark each item that is answered incorrectly. With Multiple-choice or True-False items, draw a coloured line through the **correct** answer space of the missed items. This will allow the student to know the items he missed and the correct options. Of course, each paper should be scanned prior to scoring to make sure that only one option has been chosen for each item. Any item for which the student has marked more than one answer should be scored as incorrect.
4. When separate answer sheets are used for a final exam that will not be returned to students, it is convenient to use a clear plastic sheet in the shape of the answer sheet for developing a key. The correct answers on the teacher's completed sheet are circled with a felt marking pencil on the plastic overlay. To score, the plastic overlay is placed on the student's sheet and the number of answers appearing within the circles are counted to obtain the total score. Scanning for multiple marking of an item can be done while scoring with a plastic overlay.
5. Answer sheets should only be used with children beyond the grade two level and only after the teacher is convinced that the students can handle the procedure effectively.

(b) Scoring An Essay Test: Scoring an essay test acutally begins with a clearly worded test question based upon learning outcomes expressed in behavioral terms. If the problem presented to students is vague, consistent scoring is virtually impossible. The second requirement is for the teacher to have a clear idea of what constitutes a "model" answer. At the time of composing an essay question, the teacher should make an outline containing minimum points required for a satisfactory answer.

Different approaches used in scoring essays are dependent upon the purpose of the question, the length of response and the complexity of the answer. Short answers (restricted responses) are usually scored by what may be called the "point" or "analytical" method. With this method, an answer is judged in relation to a detailed scoring key and given a number of points to indicate its degree of comparability to the ideal answer. The scoring key is usually prepared when the question is written, and this key is applied consistently to all papers.

Longer answers (extended responses of a few pages in length) may be scored by what is variously called the global or holistic method. Each answer is read and assigned to one of perhaps five piles based upon the overall quality of the response. Pile 3 would include papers of average quality, while piles 1 and 2 are reserved for below average responses and piles 4 and 5 for those of above average quality. It is recommended that each response be reread quickly at least once so that those found to have been misclassified may be reassigned.

Scoring essays present unique problems for the teacher. Teachers are also urged to refer to either the Elementary or the Secondary packages of the document entitled **Teaching and Evaluating Student Writing: A Resource Book (1978)** published by the Learning Assessment Branch of the Ministry of Education. These volumes provide writing exercises for over 40 separate writing skills, detailed procedures for their evaluation, as well as samples of student writing at different levels of proficiency.

In addition to the suggestions previously discussed, the following guidelines should help increase the precision as well as the validity of the measurement process:

1. Score all students' answers to one question before going on to the next. This procedure allows a consistent standard to be maintained, making it easier both to keep in mind the basis for judging each answer and to identify answers of varying degrees of correctness. If possible, it is recommended that the marker try to score all responses to a particular question without interruption. However, one must also be conscious of the fatigue factor in marking essay questions and not let it affect the consistency of marking standards.

2. Shuffle the papers between the grading of different questions. This procedure avoids the problem of having a particular student's paper always scored first, last, in the middle, or just before or after some talented or inept student.

3. Score the students' responses anonymously. Unless one is extremely objective, the score assigned may be unfairly biased by knowledge of a student's previous performance or other characteristics rather than by his actual response to this item. A teacher can avoid attaching a name to a particular paper by having students put their names on the back of their papers. When the identity of a paper is known, one must make a conscious effort to eliminate any bias in judgement.

4. If the papers are to be returned to students, write comments and indicate errors on the answer sheets. This is especially important for formative evaluations.

CHAPTER 3
THE ANALYSIS OF CLASSROOM TESTS

Unfortunately, many good tests are discarded and forgotten after they have been marked and the scores entered in a record book. This chapter is concerned with the very important task of analyzing and improving classroom tests. A procedure is explained that will allow both the teacher and the class to share the job of test analysis. The concepts of item difficulty and discrimination are also considered with reference to end-of-unit tests.

3.1 ANALYZING TEST ITEMS

Once a test is administered and scored, it is usually desirable to evaluate the effectiveness of each of the test items to do the job they were designed for — that is, to consistently distinguish between good and poor performers. A detailed study of how the students responded to each item can reveal areas in which construction was especially good or especially poor. It will also help to identify individual students' areas of weakness that may be in need of remediation. Although such information cannot serve to improve the items on the current test, it can form the basis for worthwhile item revision prior to reuse. This information is often found to be instrumental in improving the ability to construct tests and examinations in the future.

There are two main parts to an item analysis. First is an examination of the difficulty level of items (the proportion of students who answer each item correctly). Second is the calculation of the discrimination index of each item. This index summarizes information as to whether students who are knowledgeable in the subject matter of the test actually answered an item correctly more often than students who did not know the subject matter.

3.2 A CLASSROOM PROCEDURE FOR CONDUCTING ITEM ANALYSES

There are many ways of conducting an item analysis depending upon whether the test is norm-referenced or criterion-referenced and upon whether the teacher has help during the procedure from either students or a computer. The following steps in the "show of hands" approach are the most practical for use with the typical norm-referenced classroom test and can be performed with students at the intermediate grades and higher. In this discussion it is assumed that the analysis data will be obtained from answer sheets although, if students answer directly on the test booklet, the same procedures may be used. This procedure can be used with any questions that can be scored as either correct or incorrect.

1. **Arrange the answer sheets.** After the answer sheets have been scored, return them to the students for a quick re-checking and then recollect and arrange them in descending order, highest to lowest score on the test.
2. **Divide the answer sheets into high scoring and low scoring groups of students.** This is done by counting down to the middle of the pile and dividing the papers into two equal groups. If there is an uneven number of sheets, discard one that has a score equal to the middle score. If a number of students tie at the middle score, randomly assign an equal number of answer sheets to the high and low scoring groups. Note: there should be the same number of answer sheets in both the high and low scoring sections.
3. **Distribute the two groups of answer sheets to the class.** Pass out the papers in the high group to the students on the left hand side of the room. In order to maintain the confidentiality of an individual's score, each student should be given a code number which can be placed on the answer sheet in front of his name. If the name is on the extreme right hand upper side of the page, it can be removed with a paper cutter prior to the item analysis while the code number remains for identification purposes.
4. **Choose a student helper.** If there is an odd number of students, there will be someone without a paper who can count the "show of hands". If there is an even number of students, the teacher can choose a student to act as helper and allow another capable student to work with two answer sheets.

5. **Count correct responses from a "show of hands" and record the data.** Once the answer sheets have been distributed, explain to the students that you are going to find out helpful information about their learning which will aid them in the review and interpretation of their results. This is done by checking the total number of correct answers within the class for each question. The procedure starts with the teacher calling out each item number in turn and having students raise their hands if they hold a paper that has the particular item correct. The students holding high scoring papers first raise their hands until the helper calls out the number of students in the high group (H). Then students holding low scoring papers raise their hands and the number of papers in this group that had the particular item marked correct (L) is recorded. The teacher should record the H and L values on the test booklet to the right of each correct answer. Sometimes it adds to the interest of the class if the second helper also records the H's and L's on the board so that the students have an idea of the relative difficulty level of the various items. Once a class has become familiar with the procedure, they can complete the tallying portion of a fifty item test in approximately fifteen minutes.

Two other steps are necessary in order to prepare the item analysis information for practical use. FIRST, the teacher should add the H and L values for each item to get an impression of their difficulty values. Actually, the difficulty index can be expressed as the proportion of students in the class as a whole who answered the item correctly (p-value). It is obtained by first adding H + L and then dividing the sum by the number of student papers that were used in the analysis (N).

$$\text{p-value} = \frac{H + L}{N} = \text{proportion of students who correctly answered the question.}$$

P-values can range from 0.0 when no one in the class chose the correct answer to 1.0 where everyone in the class chose the correct answer. Note that the higher the proportion, the easier the item. (See example following).

SECOND, the discrimination index (DISC) of each item is calculated by subtracting L from H and dividing by one half the number of student papers used in the analysis.

$$\text{DISC} = \frac{H - L}{N/2}$$

This index varies from −1.0 to +1.0. A value close to +1 indicates that a test question does a very good job of distinguishing between high achieving and low achieving students. (See example following).

Example:

Subject: **Grade 3 Social Studies** Topic: **The Jungle**
Cognitive Level: **Knowledge**

Item #5

What season do we have north of the Equator when the sun is shining directly over the Tropic of Cancer?

A Winter
B Spring
C Summer
*D Fall

Calculations: for p-value and discrimination index.

1. Number of students in class (N) = 30

2. Arrange the test papers from highest scoring to lowest scoring. Divide the papers into two groups (high scoring and low scoring).

3. Count number of students from high group and from low group who answered each question correctly. Suppose for this example that 11 students from the high group, and 8 students from the low group correctly answered the question. So,

H = 11
L = 8

4. p-value $= \frac{H + L}{N} = \frac{11 + 8}{30} = \frac{19}{30} = .63$ = DIFFICULTY INDEX

DISC $= \frac{H - L}{N/2} = \frac{11 - 8}{30/2} = \frac{3}{15} = .20$ = DISCRIMINATION INDEX

3.3 INTERPRETING DIFFICULTY INDICES

The difficulty of a test is dependent on the average difficulty of the items. If a test contains items all having high p-values, then the average of the total scores for the class will also be high.

When developing criterion-referenced tests that are used in conjunction with mastery learning and individualized instruction programs, one would expect that the items would be relatively easy for the student, provided the instructional program is effective. In such a case the p-values of the items would be .80 or higher; that is, 80% of students would correctly answer each item.

The difficulty level of items in a norm-referenced test should vary depending upon the purpose of the test. If the purpose is to select a few high ability students, then the average p-value of the items should be low. That is, only the very able students in the class will be answering the questions correctly. However, if the purpose is to select low ability students, the average p-value should be relatively high because only the least able students will be having difficulty with the question.

One never knows prior to testing exactly what the difficulty index (p-value) of an item will be. The teacher can only guess at the approximate value when building a test and should plan to have the p-values in tests used for grading purposes to range from approximately .30 to .70.

One must observe some guidelines in interpreting difficulty indices. A high p-value for an item may not necessarily mean that the students actually know the subject matter of the item. The item may have been easy because of a structural defect such as a grammatical clue. If the students noticed the clue, perhaps they responded correctly to the item without knowing the answer. Items with low p-values might be hard for a number of reasons. The key may have been incorrect for that item and should be checked. Another possibility is that more than one correct answer was possible. The wording of such questions should be given a close examination.

It should be kept in mind that the difficulty index for items is not an absolute value, but is indicative only of the relative difficulty of the item with a particular group of students at a particular point in time. A somewhat different group of students or the same students several weeks earlier or later may have responded to the item quite differently. In statistical terms, this difficulty index is "sample dependent". By keeping a regular record of the difficulty index each time the item is used with different groups of students, as was shown in Figure 2.1, a better appraisal of the way the item "works" with students can be acquired.

It may be of interest to the reader to know that considerable research has been conducted in the past five years to arrive at a sound procedure for determining item difficulty that is not sample dependent. One alternative approach is referred to as "Latent-Trait Analysis". The 'Rasch' method, used extensively in Oregon and more recently by the B.C. Learning Assessment Branch, is one of several forms of latent-trait analysis. This statistical procedure makes it possible to determine the difficulty indices of large banks of test items on a single continuous scale (a B.C. Mathematics test item bank will include over 2500 items for grades 3, 4, 7, 8, and 10) even though any one student would not have been given more than 30 or 40 of the questions in that bank. More importantly, this method provides stable estimates of item difficulty across different samples of students and thus overcomes the problem of sample-dependence. Further information on these developments will be forthcoming in a separate document as the tests become available.

3.4 INTERPRETING DISCRIMINATION INDICES

In norm-referenced measurement, the purpose of a test is to measure individual differences within a class. In such tests the discrimination indices should all be positive and as high as possible. There are two ways to study the discrimination indices of items.

First, calculate the H minus the L value of the items. That is, subtract the number of students in the Low group that got the item right from the number in the High group who did so. For good items the difference between these two values should be equivalent to at least 10% of the class. For example, in a class of 36 students at least 4 more students in the High group than in the Low should get the item correct. When items are very easy or very difficult, this tends to handicap them from being good discriminators and the standard may be lowered from 10% to 5% of the class. In general, extremely difficult or extremely easy items will show very little discrimination. However, some items of this type are often necessary in order to have adequate and representative sampling of the course content and objectives.

A second approach is to actually calculate the discrimination indices using a formula $DISC = \frac{H - L}{N/2}$ rather than use the H minus L values, and then interpret the indices according to the following criteria:

If the discrimination index is:	Judge the item as:
.40 or higher	very good
.20 - .39	satisfactory
below .20	poor, reject or revise the item

It should be noted that the criteria for judging discrimination indices although useful as a guide, should not be followed too rigidly. Items statistics, as noted above, are "sample dependent" and do tend to fluctuate from one group of students to another because of the small numbers of students typically involved in classroom item analyses. In contrast, test publishing companies usually calculate the items statistics for their tests after administering them to at least 400 people. The item analysis data will also vary depending upon the level of ability of the students, their educational background and the type of instruction they have had.

A high positive discrimination index suggests that the item is measuring the same general factors that the test as a whole is measuring. A low discrimination index does not necessarily indicate that the item is defective, however. For example, if an item is measuring an important content area that is considerably different from the majority of other items on the test, it still might be a good item even though the discrimination index turns out to be quite low.

When a discrimination index turns out to be negative, the item must be studied carefully to see why the better students have more trouble with it than weaker students. Sometimes asking the students why they answered the way they did will reveal a flaw in the item construction, or may suggest alternative procedures for teaching that concept. Of course, an item with a large negative discrimination index should be checked to make sure it was not keyed incorrectly or does not have more than one correct answer.

3.5 ITEM ANALYSIS BY A TEACHER

When it is either inappropriate or inconvenient for the class to take part in an item analysis, one alternative is for the teacher to perform the analysis.

With only one or two classes, it is convenient to choose the upper and lower groups by counting off the top ten and bottom ten answer sheet. The remaining sheets are placed aside and not used in the analysis. For more than two classes, one can use the top one third and bottom one third of the total number of students.

The procedure that can be used with multiple-choice questions follows. First, arrange the top ten answer sheets so that they are overlapping and just the 'A' response column on each sheet is visible. Then, with the answer key placed on the bottom of the answer sheet pile, count and record the H values for all items where the correct response is given as A on the key. Repeat this process with all other letter columns, and then repeat the entire process with the low scoring answer sheets. Recording can be done directly on the test item cards.

Interpretation of the difficulty indices is the same as with the show-of-hands method described above. Of course, with the teacher method, the divisor used for calculating the difficulty index is the number of student papers used for the analysis, not the number of students in the class. The standards for judging the discrimination index have to be modified slightly due to the fact that a number of cases in the middle of the score distribution have been left out of the analysis. When the top and bottom 10 sheets are used, H – L should equal 3 or more for items with difficulty indices between .30 and .70. This difference between H and L can be lowered to 2 for items with extreme difficulty indices.

Example:

Subject: **Grade 3 Social Studies** Topic: **The Jungle**
Cognitive Level: **Knowledge**

Item #5

What season do we have north of the Equator when the sun is shining directly over the Tropic of Cancer?

A Winter
B Spring
C Summer
*D Fall

Calculations: H = 4, L = 1

$$p = \frac{H + L}{20} = \frac{5}{20} = .25$$

$$DISC = \frac{H - L}{N/2} = \frac{3}{10} = .30$$

Item #5	A	B	C	D	OMITS	P	Disc
Upper N = 10	1	3	1	4	1	.25	.30
Lower N = 10	4	2	1	1	2		

Comments:
Very difficult. Review in detail.

CHAPTER 4
SUMMARIZING AND INTERPRETING TEST PERFORMANCE

The purpose of this chapter is to provide minimum knowledge about and skills in using elementary statistical techniques so that the planning, use and evaluation of teacher-made tests may be facilitated.

Briefly the chapter is divided into four major parts. The beginning sections deal with how test scores can be organized and described. Following that are sections dealing with the interpretation of test scores, including the concepts of reliability and measurement error.

The chapter advocates the use of simple, practical and short-cut procedures for handling numerical test data. Although this approach is subject to some error, this is not serious enough to weaken its usefulness in analyzing classroom tests.

Table 4.1

Three Sets of Test Scores for 30 Pupils

Pupil	Test #1	Test #2	Test #3
A	24	17	25
B	30	30	30
C	32	23	27
D	36	26	39
E	26	17	23
F	22	14	19
G	36	24	27
H	30	21	33
I	34	24	30
J	37	22	37
K	22	19	25
L	33	31	32
M	28	20	25
N	33	19	32
O	38	31	35
P	37	29	37
Q	30	29	29
R	31	20	26
S	32	25	34
T	31	25	32
U	31	25	31
V	37	25	34
W	32	29	40
X	37	24	39
Y	23	22	30
Z	28	25	25
AA	28	22	33
BB	23	20	25
CC	32	16	26
DD	31	25	21
Maximum score Possible	45	42	50

4.1 ORGANIZING AND DESCRIBING A SET OF TEST SCORES

(a) Constructing Frequency Distributions: Table 4.1 contains a list of the raw score obtained by 30 students on three different tests. When the scores are presented in this form, similar to a teacher's class record book, it is difficult to make any generalizations regarding the ranges, from highest to lowest score, the averages, etc. These scores can also be arranged into frequency distributions, as shown in Table 4.2. When this is done all possible scores are listed in order of size and the frequency (f) of students that obtained each score is provided.

A frequency table makes it possible to observe some characteristics of test performance that were not obvious before. For example, it is immediately obvious that the highest score on the Test 1 is 38, the lowest 22; nearly one half of the class obtained three scores, namely 37, 32 and 31; no one in the class obtained scores of 35, 29, 27 or 25. Comparing across different distributions it can be seen that the range is about the same for each test, where the range is defined as the number of possible scores between and including the highest and lowest scores. (Range = Highest score minus lowest score plus one).

Notice that the frequency distributions in Table 4.2 allow for rather crude norm-referenced interpretation of an individual's score to be made. For example, it can be seen that Pupil A's three test scores from Table 4.1 (24, 17, 25) each fall in approximately the lowest quarter of their respective distributions on Table 4.2 Score 24 is the fifth from the bottom, score 17 is fifth from the bottom and score 24 is eighth from the bottom.

(b) Graphing Frequency Distributions: Sometimes it is useful to present a frequency distribution graphically rather than in tabular form. Graphical presentations are of value particularly when presenting qualitative data to parents. The histogram in Figure 4.1 presents a visual picture of the scores of Test 1 after they had been organized into the grouped frequency distribution shown in Table 4.3.

Table 4.2 Frequency (f) Distributions for the Three Sets of Test Scores Contained in Table 4.1[1]

Test 1		Test 2		Test 3	
Score	f	Score	f	Score	f
38	1	31	2	40	1
37	4	30	1	39	2
36	2	29	3	38	0
35	0	28	0	37	2
34	1	27	0	36	0
33	2	26	1	35	1
32	4	25	6	34	2
31	4	24	3	33	2
30	3	23	1	32	3
29	0	22	3	31	1
28	3	21	1	30	3
27	0	20	3	29	1
26	1	19	2	28	0
25	0	18	0	27	2
24	1	17	2	26	2
23	2	16	1	25	5
22	2	15	0	24	0
		14	1	23	1
Range = 17				22	0
		Range = 18		21	1
[1]For example, in Test 1, one student had a score of 38, 4 students scored 37, and so on.				20	0
				19	1
				Range = 22	

Figure 4.1 Histogram Based on Test #1 Scores of 30 Students

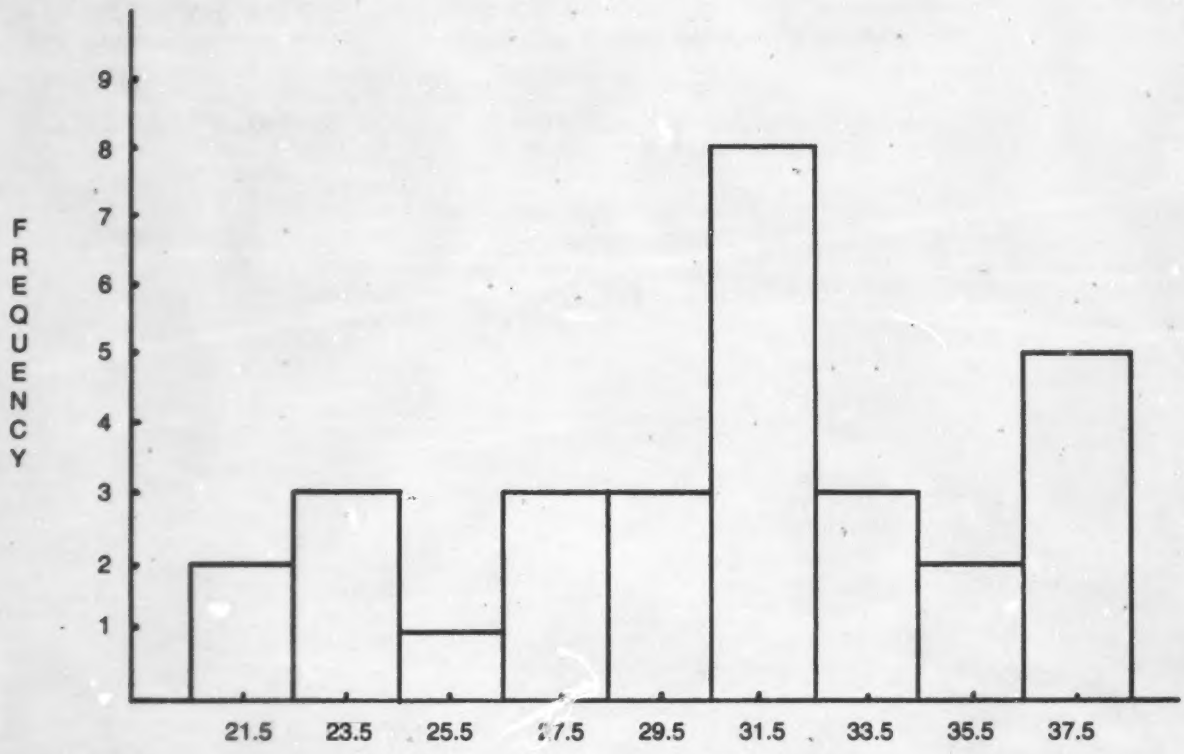


Table 4.3 Grouped Frequency Distribution of Test #1 Scores

Score Interval	Frequency
37-38	5
35-36	2
33-34	3
31-32	8
29-30	3
27-28	3
25-26	1
23-24	3
21-22	2

In developing a grouped frequency distribution, the scores are combined into intervals of predetermined and uniform size (usually two, three or five points on the score scale) so that the graph will contain ten to fifteen groupings. In Figure 4.1, the size of each interval is two and, as such, the scores are classified into nine different columns. The mid-point of each interval is printed along the horizontal axis, while the corresponding frequencies (numbers of students) are read from the vertical axis.

Grouping scores is particularly desirable when the class is large or when the range of scores is great.

4.2 MEASURES OF CENTRAL TENDENCY

One of the ways we can describe a distribution of scores is to determine its central tendency, i.e., the score around which the distribution tends to centre or the score that describes the general level of performance of the class. There are two commonly used measures of central tendency; the mean (the arithmetic average) and the median.

(a) The Mean: To calculate the mean one needs to know the exact value of each test score. All the scores are then added and the sum divided by the number of students.

$$\text{Mean} = \frac{\text{Sum of all the Scores}}{\text{Number of Students}}$$

The mean score, calculated to one decimal place, for the three distributions in Table 4.1 are as follows:

$$\text{Test \#1} \quad \text{Mean} = \frac{924}{30} = 30.8$$

$$\text{Test \#2} \quad \text{Mean} = \frac{699}{30} = 23.3$$

$$\text{Test \#3} \quad \text{Mean} = \frac{901}{30} = 30.0$$

It should be mentioned that if a few scores are either much higher or much lower than the majority of other scores, then the mean is pulled in the direction of these deviate or extreme scores.

(b) The Median: The median, in contrast to the mean, is not dependent on the value of each score in the distribution. Extreme scores have no effect on the median. The median is the score point in the distribution that divides it into two equal parts based on the frequencies of the various scores. For practical purposes it is not important to calculate the exact point. The actual score closest to the point, called the **discrete median,** is appropriate for most classroom purposes. The discrete median for each of the frequency distributions in Table 4.2 are listed below.

Test #1 Discrete median = 31
Test #2 Discrete median = 24
Test #3 Discrete median = 30

For Test 1 it can be seen from Table 4.2 that score 31 is about at the centre of the distributic . Note that there are 14 scores above and 12 scores below 31. Procedures used to calculate the exact median are described in most measurement texts listed in the References.

Notice in Table 4.2 that in each of the three distributions the scores are about equally divided above and below the means and medians for each of the tests. This is because the two measures of central tendency are approximately equal within each distribution. In other situations, called **skewed distributions,** when there are a large number of scores at one end of the distribution and only a few cases spread out at the other end, the mean and median will not be close together. The mean will tend to be closer than the median to the end of the distribution that has scores with small frequency values because it is influenced by these extreme scores.

(c) Uses of the Mean and Median:

1. The mean is a useful measure when the distribution is symmetrical about the centre or when a teacher wants extreme scores to play a significant role in determining central tendency.
2. The mean must be calculated when raw scores are to be changed to standard scores* or when further statistical techniques are to be employed. It "goes with" the standard deviation and is basic to the calculation of correlation coefficients*.
3. The mean is used when the central tendency of different groups are compared on the same test. It is more stable, consistent and reliable than the median.
4. The median is an appropriate measure when the distribution is markedly skewed. That is, when there are extreme scores that make the mean unreliable. For example, suppose five teachers have the following salaries:

$12,600
$12,700
$12,800 ——— Median
$12,900 ——— Mean $13,200
$15,000

The median ($12,800) is more appropriate than the mean ($13,200) as a measure of central tendency as it is more representative of the majority of teachers.

5. The median is appropriate when a teacher needs a quick average for reporting purposes.
6. If the numerical data is converted to ranks, the middle individual would be placed at the median on the attribute measured.
7. When distributions are approximately normal (such as those obtained from most standardized tests), it makes little difference which measure of central tendency is used.

4.3 A MEASURE OF SCORE VARIABILITY

The mean and the median give some idea of the value of the average score in a distribution. However, most decisions, particularly those dealing with norm-referenced test interpretations are concerned with the extent of individual differences within a group. A statistic, more precise than the range, that reports the amount of variability or of how "spread out" the scores within a distribution are, is called the **standard deviation.** It measures the degree to which scores deviate from the mean of the distribution.

(a) The Standard Deviation: Exact procedures for determining the standard deviation of a distribution are easy to employ provided one has a hand calculator, otherwise the procedure is rather tedious. A simple method for estimating the standard deviation of a distribution of scores (SD) is given by the formula below, appropriate for use with most norm-referenced classroom distributions.

$$SD = \frac{\text{(Sum of highest 1/6 of scores) minus (sum of lowest 1/6 of scores)}}{\text{one-half the total number of scores}}$$

* The reader is referred to the Glossary at the end of the Manual for some detail on these terms.

The standard deviation, rounded to one decimal place, for each distribution in Table 4.2, is calculated as:

$$\text{Test \#1} \quad SD = \frac{(38+37+37+37+37) - (24+23+23+22+22)}{15}$$

$$= \frac{186-114}{15} = \frac{72}{15} = 4.80 \qquad = 4.8$$

$$\text{Test \#2} \quad SD = \frac{(31+31+30+29+29) - (19+17+17+16+14)}{15}$$

$$= \frac{150-83}{15} = \frac{67}{15} = 4.46 \qquad = 4.5$$

$$\text{Test \#3} \quad SD = \frac{(40+39+39+37+37) - (25+25+23+21+19)}{15}$$

$$= \frac{192-113}{15} = \frac{79}{15} = 5.26 \qquad = 5.3$$

The three standard deviations of 4.8, 4.5 and 5.3 show that each test measures individual differences within the class to about the same extent. That is, on the average, the scores spread out from the mean a distance of approximately five raw score points in each distribution.

(b) Uses of the Standard Deviation: The standard deviation is used extensively in statistics and educational measurement. Listed below are some of the more common uses. It is not expected that an individual who is new to the field of education measurement will fully understand all of the concepts involved. Some of the uses will be dealt with later on in this booklet. These and other uses can be studied in more detail in the texts listed at the back of this booklet.

1. Standard deviation units include a constant percentage of frequencies within different sections in a normal curve. This relationship forms the basis for making statistical inferences in educational research. It is also used in the interpretation of normalized standard scores.

2. The standard deviation is required for calculating the extent to which variable errors are associated with test scores. A basic concept called the **standard error of measurement** is used for interpreting scores on both teacher-made and standardized tests.

3. The standard deviation is the basic measurement unit associated with **linear standard scores** (Z). These scores are used for interpreting nearly all standardized achievement and ability tests.

4. The standard deviation can be used to compare the extent of variability or the degree of homogeneity within different sections of the same course or classes of the same grade level.

5. The standard deviation can be used as a basis for weighting different tests when two or more test scores are added to form a composite score.

4.4 INTERPRETING TEST SCORES

Interpreting test results starts first with the determination of each individual's raw score; that is, the number of right answers obtained on the test. Although raw scores are useful for describing certain characteristics of a class, raw scores are uninterpretable in themselves especially when individual students are to be studied. For example, having no other information, what can be said about Ann's raw score of 37? This score must be set in terms of either a criterion-referenced or norm-referenced framework in order to have meaning.

A score obtained from a criterion-referenced test is interpreted in terms of an individual's status with respect to a well-defined instructional objective. Ideally, test publishing companies will provide teachers with technical manuals to aid them in this type of score interpretation. Scores from teacher-made tests and other norm-referenced tests are interpreted after first converting raw scores into other derived scores such as stanines or percentile ranks. The procedures for converting raw scores into various types of derived scores and other statistical procedures helpful for test interpretation are described in many of the texts listed in the bibliography.

Table 4.4 provides a list of questions one might ask about a set of test scores and suggests the types of scores that are most useful for answering these questions.

Table 4.4

Interpreting Tests Using Different Types of Scores

Questions about the interpretation of test results	Answer the question using the following types of scores:
1. What was the highest possible score?	Raw score
2. What was the highest score and the lowest score actually obtained by students?	Raw scores
3. What was the average score obtained by the class, the grade, or the district?	Raw scores used to calculate the mean
4. Ann received a score of 26 on test X. What percentage of students in the class, grade or district scored lower? What percent scored higher?	Percentile Rank*
5. Was Ann's test score of 26 on Test X any better than her score of 20 on Test Y?	Percentile Rank* or Stanine
6. Which Test, X or Y, was the more difficult?	Average raw score on each test converted to percent
7. On which test, X or Y, was variation among students' scores the greater?	Raw scores are used to calculate the standard deviation.
8. Were the test scores spread symmetrically and smoothly or skewed and unevenly?	Raw scores are used to plot a histogram. (graph)
9. Is there a relationship between how well students did on Test X and Y?	Raw scores are used to calculate the correlation coefficient.
10. Which Test, X or Y, is the more reliable (i.e., internally consistent)?	Raw scores used to calculate reliability coefficient.

*The reader is referred to the Glossary at the end of the book for some detail on these terms.

4.5 RELIABILITY

The reliability of a set of test scores refers to the degree to which score differences are actually **dependable** and **stable** estimates of the students' mastery of the material being tested as opposed to being the result of chance or random factors. Reliable test data may not be valid. Consistency in measurement (reliability) does not necessarily equate with **truthfulness, value** or **worthwhileness** (validity). Highly reliable test data are not a guarantee that a test is valid. Low reliability, however, particularly with norm-referenced measurement, would indicate that the data are invalid for making any type of educational decision. Thus, a teacher must check both the validity and reliability of a test in order to use the results with assurance.

Reliability can be studied from two separate but related points of view — as a mathematical theory of test scores and as a practical problem of test construction and interpretation. This section will deal only with the latter aspect. The theoretical approach to the concept of reliability may be found in various references at the end of the chapter.

First, two simple and related procedures will be presented which allow a teacher to estimate the reliability of test data. Next, the interpretation of reliability through the use of standard error of measurement will be considered and finally suggestions will be given to improve the reliability of classroom tests.

(a) Methods of Calculating a Reliability Coefficient: A number of techniques are available for calculating a reliability coefficient (similar to a correlation coefficient that varies from 0.0 to 1.0). All of these techniques are based on the concept of correlation (see the Glossary). One method involves two administrations of a test to the same students. Another method requires different forms of a test to be administered to the same students. In the latter case, various interpretations are placed on the reliability coefficient depending on the time interval between the test administrations. The foregoing procedures for estimating reliability are most practical for standardized tests and, therefore, will not be dealt with in this handbook. For teacher-made tests a number of simple and practical techniques, generally called **internal consistency** or **homogeneity,** are available for estimating reliability. Two methods, Kuder-Richardson and Saupe, will be illustrated using the statistics obtained from the thirty scores of Test #1 presented previously in Table 4.1.

Kuder-Richardson Technique: Kuder and Richardson developed a number of formulae for estimating reliability. Their formula Number 20 (KR-20) is used extensively with standardized tests. Their formula Number 21 (KR-21) is appropriate for teacher-made tests:

$$\text{KR-21} = 1 - \frac{\bar{X}\,(k-\bar{X})}{k(SD)^2}$$

where
k = the number of items on the test
$\bar{X}$ = the mean of the test scores
SD = the standard deviation of the test scores
SD^2 = the square of the standard deviation

Substituting values for Test #1, discussed earlier, yields the following:

$$\text{KR-21} = 1 - \frac{30.8\,(45-30.8)}{45\,(4.8)^2}$$

$$= 1 - \frac{(30.8)\,(14.2)}{(45)\,(4.8)\,(4.8)}$$

$$= .58$$

Thus the reliability coefficient for Test #1, using the KR-21 technique, is .58. Before interpreting this value, the Saupe method will be presented and then the two estimates will be interpreted and compared.

Saupe Reliability: J. L. Saupe (1961) developed and even simpler reliability formula than the KR-21. Actually, it is an estimate of KR-20. The formula is:

$$\text{Saupe Reliability} = 1 - \frac{.19k}{(SD)^2}$$

where k = the number of items on the test
SD^2 = the square of the standard deviation.

For Test #1 with k = 45 and SD = 4.8 the reliability calculation is as follows:

$$\text{Saupe Reliability} = 1 - \frac{.19\ (45)}{(4.8)^2}$$

$$= 1 - \frac{8.55}{23.04}$$

$$= .63$$

Using the Saupe formula provides a reliability estimate slightly higher than the KR-21 procedure. However, from a practical point of view, the two estimates are quite comparable. In general, these methods will usually produce an underestimate of the "true" internal consistency coefficient (see Glossary).

(b) Interpreting a Reliability Coefficient: As stated earlier, the KR-21 and Saupe methods estimate the internal consistency of a particular test given to a particular group at a particular time. If any of the above factors (test, group, time) were changed, the resultant coefficient would likely change. As such the coefficient should not be considered a characteristic of the test. Rather, it is an estimate, from the test score distribution and test length, of the degree to which pupils who obtained high test scores on one set of items on the test also obtained high scores on other sets of similar items. Technically, internal consistency reliability is a measure of the homogeneity of the various items on the test.

Reliability coefficients range from 0.0 to 1.0. The closer the coefficient is to 1.0 the more confidence one can have in the usefulness of the test data for making decisions. Generally, important decisions concerning individuals should not be made unless the reliability coefficient is .90 or higher. However, when comparing differences between groups (or classes), data that yield correlation coefficients of .60 or higher would be considered satisfactory. A well constructed objective classroom test could yield a reliability coefficient of at least .60, whereas the reliability coefficients of standardized test batteries are usually greater than .90. (Using these standards the reliability coefficient for Test #1 could be considered barely within the acceptable range). Combining the scores of three or more well constructed classroom tests would likely raise the reliability of the resultant composite total to a level that would be acceptable for making decisions about individual students.

When interpreting Kuder-Richardson and other internal-consistency correlation coefficients one must be sure that the following assumptions are reasonably met:

1. The formulae should be used only with objectively scored tests in which each item is scored 1.0 (correct answer) and 0.0 (incorrect or omitted answer). Essay tests, where items may have variable credit, require an analysis of variance procedure for estimating reliability (see Ebel, [1972] pp. 419-420). However, due to the complicated mathematics involved and the extensive time needed for hand calculation, such a procedure is not practical for classroom use.
2. In general, only one type of item should be used. Do not, for instance, mix true-false and multiple-choice items in a single reliability calculation.
3. All items should be measuring the same characteristic (trait). A test measuring a great many intellectual skill and cognitive levels as well as measuring widely divergent content areas will produce an internal consistency reliability coefficient that is seriously low and, hence, inappropriate.

4. Internal consistency reliabilities should be computed for power tests only — that is, tests where most students have sufficient time to finish. To the extent that speed plays a part in determining response, internal consistency methods will produce spuriously high coefficients.
The reliabilty of speed tests (produced by some standardized test companies) must be estimated by using procedures other than internal consistency methods. These procedures are treated in detail in many of the texts listed in the references.
5. The KR(21) is appropriate only for tests that have a rather narrow range of medium sized difficulty indices (i.e., between .30 and .70).
6. Of course, reliability coefficients refer only to the specific group who wrote the test. Generalizing across groups is not appropriate unless they are quite similar in academic background and other characteristics related to test performance.

The most useful and practical way of interpreting the reliability of test data is through the use of the **standard error of measurement.** The following section explains that concept and its application.

4.6 STANDARD ERROR OF MEASUREMENT

Previous sections have described how to estimate the reliability coefficient for test data based on how the total group (or class) responded to the various test items. The reliability coefficient estimates the accuracy of the measurement results as a whole. The **standard error of measurement,** however, permits one to interpret a reliability coefficient in terms of the accuracy of an individual's score.

It is highly unlikely that one administration of a paper and pencil test will measure an individual's "true" level of achievement or ability. Various factors combine to produce error in the measuring process. Among these factors are a student's health on the day of the testing, emotional condition, motivation, rapport with the teacher, recent practice in the subject matter tested, luck in guessing, as well as fluctuations in attention, memory and fatigue. Other factors within the test such as inadequate or limited sampling of content will also cause error in the test results.

The standard error of measurement estimates the amount of **random error** associated with each student's score and expresses this amount in terms of score units. The random error is assumed to be normally distributed around each score and the standard error of measurement is an estimate of the standard deviation of the random error distribution. The formula is:

$$SEM = SD\sqrt{1-r} \qquad \text{where}$$

SEM = the standard error of measurement
SD = the standard deviation of the test scores
$\sqrt{\ }$ = "take the square root of"
r = the reliability coefficient of the test data

Referring to Test #1, which has a Saupe reliability coefficient of .63 and a standard deviation of 4.8 and after substituting these values in the formula, we have

$$SEM = 4.8\ \sqrt{1-.63} = 2.9 \text{ or approximately } 3.0$$

Theoretical interpretations of the standard error of measurement are discussed in most educational measurement texts. However, let us consider a practical application of the concept by applying it to Student A's score of 24 on Test 1. If Student A wrote a large number of comparable forms of Test 1, his scores would vary by plus or minus one standard error of measurement (3 raw score points) about two-thirds of the time. With reference to Test 1, the standard error of measurement can be applied to Student A's score as follows:

Raw score ± Standard error of measurement = Expected range

24 ± 3 = 21 to 27

This expected range 21 – 27, called a **confidence interval,** is the extent to which one could expect Student A's score to vary on comparable tests approximately two-thirds (68%) of the time due to the unreliability of the measuring instrument.

If we wished to be more confident of Student A's "true" achievement level, we could increase the size of the confidence interval so as to include two standard errors of measurement on either side of his raw score as follows:

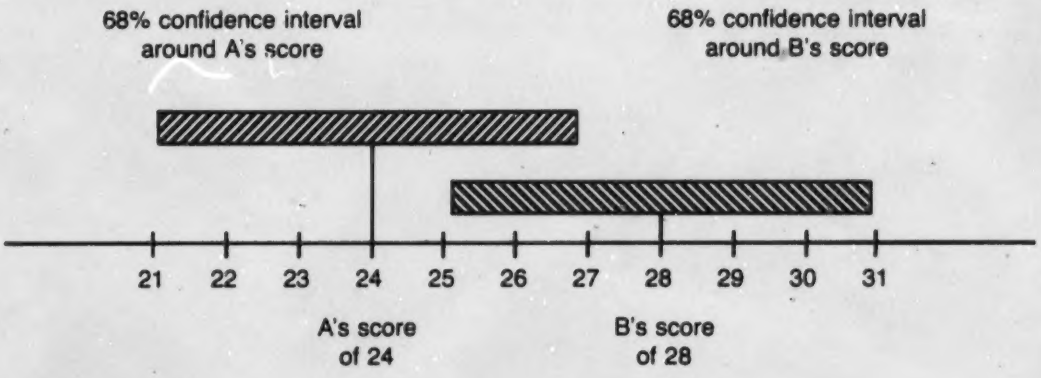
Raw score ± 2 (SEM) = 95% confidence interval

24 ± 2(3) = 18 to 30

Theoretically, if Student A was tested with comparable forms of Test #1 a large number of times, ninety-five percent of his raw scores would be expected to fall within the confidence interval 18 – 30. The remaining five percent of his scores would fall outside the interval.

In this case, we could feel quite confident that the interval (18–30) included Student A's "true" score. However, as a "true" score is a theoretical score assumed to be obtained from a perfectly reliable (and therefore non-existent) measuring instrument, we can never really know a student's "true" score. We must be content in knowing that it falls within a given score range and even then realize that our knowledge is not certain — only a best estimate.

Besides cautioning us on the extent of inaccuracies in measurement data, the use of confidence intervals facilitates making comparisons between students. For example, the difference between two students' (A and B below) scores are seen in a somewhat different light when they are presented as ranges of scores based on the standard error, rather than as two absolute scores. What first appears as a clear indication of superior performance by B is put into question when measurement error is taken into account. If Student B's true score is at the lower end of the confidence interval determined for the score of 28 while Student A's true score is at the high end of the confidence interval around the score of 24, then Student A would actually be performing at a higher level than Student B even though their raw scores suggest otherwise.



4.7 SUGGESTIONS FOR IMPROVING RELIABILITY

When planning, constructing and scoring tests, a teacher should be constantly aware of ways to improve the reliability of the test data. Several suggestions are listed on the following page.

1. **Compose a long test,** provided the long test does not decrease student motivation, increase student fatigue, or otherwise turn the test from one of power to one of speed.

The relationship between test length and reliability is expressed by the general Spearman-Brown formula (Ebel, 1972, p. 413). If one has a 10-item test with a reliability of .20, adding 10 items of the same type will increase the estimated reliability to .33. If one adds 20 items of the same type, the reliability estimate increases to .43.

In general, adding items to a test will have a greater effect when the original reliability is low rather than when it is high. The added items, of course, must not be just a repetition of the items already on the test. They must be a more representative sample of the hypothetical pool of test items and, as such, call upon the student to exhibit a wider range of behavior relevant to the course objectives. Adding poor items to a test can, in contrast, actually lower reliability.

2. **Compose items of medium difficulty.** Items that are correctly answered by all or failed by all contribute no variance to the test scores and thus reduce test reliability. (Note the importance variance plays in formulae of each of the reliability coefficients). Items that are correctly answered by about 50 percent of the students have the greatest potential for contributing toward high test reliability. There is, however, one exception to this rule. The first item should usually be a very easy one designed so that students can begin the test with some feeling of self-confidence.

3. **Choose an item that helps increase reliability.** Good multiple-choice items are usually more reliable than good true-false items on tests of equal length.

4. **Make sure all items are worded properly with the use of appropriate vocabulary and grammar.** Following the suggestions for writing test items (see pp. 15-20) will help improve the reliability of classroom tests.

5. **Increase the objectivity of scoring procedures.** A carefully prescribed set of standards and procedures used in scoring will tend to insure minimum reliability of test data, particularly when marking essays or themes.

CHAPTER 5
PROCEDURES FOR SETTING STANDARDS

To many people, test scores have inherent meaning. Many feel that a score of 50%, for example, represents the cut-off between passing and failing on all tests. Yet we know that a score of 50% on a complex problem-solving test in mathematics might represent outstanding achievement for a grade 4 student. Similarly, if we expect grade 10 students to achieve mastery in basic consumer math skills, a score of less than 75% on a test might be considered unacceptable for these students. Unlike physical measurements, educational measurements (i.e., test scores) are meaningful when used in some kind of comparison or when interpreted against some kind of standard. But we know that educational standards are not always understood or easily defined.

The test scores displayed earlier in Table 4.1 and 4.2 provided us with considerable information. However, the scores don't tell us what cut-off score should be used to identify those students who have mastered the material covered in the test and those who have not. The scores don't tell us what an "A" grade is on the test, or a "C", a "B", or "D". Yet these are questions that teachers must deal with on a regular basis in order to decide if Fred has successfully mastered a basic level of arithmetic, or if Maria should receive further enrichment in an English class.

Setting standards would not be a problem if students who have mastered the content measured by a test would always answer all the questions correctly and if students who have not mastered the content would get zero. In the real world of the classroom, however, we rarely get such clean-cut results.

5.1 SETTING A STANDARD ON A CLASSROOM TEST

There are many ways of setting standards on a classroom test. Two methods, which are derived from procedures suggested by the Educational Testing Service (ETS), are presented here as examples only. You may wish to modify these procedures or examine alternate methods referred to in the bibliography.

It is extremely important to realize, at this point, that all methods of setting standards depend on subjective judgment. There is simply no good method of setting standards just by plugging numbers into a formula.

The basic purpose of the methods outlined here is to identify students requiring remediation — to distinguish between "masters" and "non-masters", or between those who have "passed" and those who have "failed". The same methods could be used to define cut-off scores on a test for the purpose of assigning letter grades.

It is worth emphasizing here that there is little reason to set standards unless there is a willingness to allocate time, energy and resources to help students falling below the minimal standard, and unless one is willing to challenge all students to reach higher levels of achievement.

Figure 5.1 Teacher's Record Form

TEACHER'S RECORDING FORM

Test Title: _____

Judge's Name: _____

Question Number	Estimated Probability	Question Number	Estimated Probability
1		39	
2		40	
3		41	
4		42	
5		43	
6		44	
7		45	
8		46	
9		47	
10		48	
11		49	
12		50	
13		51	
14		52	
15		53	
16		54	
17		55	
18		56	
19		57	
20		58	
21		59	
22		60	
23		61	
24		62	
25		63	
26		64	
27		65	
28		66	
29		67	
30		68	
31		69	
32		70	
33		71	
34		72	
35		73	
36		74	
37		75	
38			SUM

(a) Procedure 1; Minimally Acceptable Performance: The following example deals with an attempt to define "minimally acceptable performance" on a test.

1. Make a copy of the Teacher's Recording Form (Figure 5.1).
2. Give each teacher a copy of the test and a copy of the Recording Form. Have your colleague look at the first question and state the probability that a minimally competent student would answer the question correctly. This task may require some time to explain. If the judges are not comfortable dealing with probabilities, ask them to think of a group of 100 **minimally competent students** and state how many of those students would be expected to answer the question correctly. Obviously, the easier the question, the higher the probability will be. The probability must be between .00 and 1.00. If the questions are multiple-choice, the probability should never be lower than the chance of guessing the correct answer by blind luck. For multiple-choice tests with four options, this probability would be .25.*
3. Have each teacher announce his or her choice of a probability for a question. Write these numbers on a blackboard or a large sheet of paper so that all the teachers can see them. Then ask the teachers who stated the largest and smallest numbers to explain the reasons for their choices. Then tell the teachers they may change their choices if they want, but not to announce them. Instead, they should simply write their revised choices on the Teacher's Recording Form.

Repeat this set of steps for each question on the test. To combine the judgements to set a standard, follow the steps below:

1. Use a hand calculator to add up the probabilities on each form. Write the sum in the box labelled SUM.
2. Now calculate the average of the sums. Simply add all the sums and divide by the number of sums (i.e., the number of teachers).

In the example which appears below, three teachers are involved in setting standards for a test made up of ten questions. To compute the standard, take the average of 5.25, 5.20, and 5.30. The sum of the three numbers is 15.75. Dividing by three gives the average of 5.25.

Figure 5.1 Examples of Three Teachers Recording Forms

Teacher's Recording Form — Teacher #1	
Question Number	Estimated Probability
1	1.00
2	.90
3	.80
4	.70
5	.35
6	.45
7	.25
8	.30
9	.25
10	.25
SUM	5.25

Teacher's Recording Form — Teacher #2	
Question Number	Estimated Probability
1	1.00
2	.85
3	.85
4	.70
5	.35
6	.40
7	.25
8	.30
9	.25
10	.25
SUM	5.20

Teacher's Recording Form — Teacher #3	
Question Number	Estimated Probability
1	.95
2	.80
3	.80
4	.65
5	.40
6	.45
7	.35
8	.35
9	.30
10	.25
SUM	5.30

For the example the standard would, therefore, fall between 5 and 6. A student scoring 5 or less would fall below the standard. A student scoring 6 or more would be above the standard.

If you wish to establish standards for a test only you are using, you may feel it is not necessary to involve other teachers in the standard-setting process. However, if the test is to be used in making important decisions about individual students, it is extremely important — given the inherent subjectivity of defining standards — to involve your colleagues on staff, or in the district. This involvement is particularly important when other teachers plan to use the same test. In short, the process used to set standards will have a great impact on the acceptability of the standards which are set.

*If you wish to define a cut-off score for an 'A' grade, you may consider "the probability that an 'A' student would answer the question correctly." The same process could be used for B and C levels, or whatever marking scheme is being used.

(b) Procedure 2; Borderline Group: This method requires a group of students whose achievement is judged to be not quite adequate, but not quite inadequate. The method is simply to identify these students and find their median test scores. Then choose this score as an estimate of the standard.

The first two steps in this procedure are identical to those given for the methods above: 1) select teachers, and 2) define minimally acceptable performance. Obviously, it is crucial that the judges be familiar with the students' levels of performance. In classes in which the objectives of instruction match the objectives measured by the test, an award of the lowest passing grade may be one indicator of minimal mastery status, but beware of the effects of variables other than student performance, such as student behavior, on the grading process.

The third step is to have each teacher submit a list of students whose performances are so close to the borderline between acceptable and unacceptable that they cannot be classified into either group.

Administer the test. When the scores are received, simply compute the median or middle score of the borderline students. That score is used as an estimate of the standard.

If the scores of the borderline-group are spread widely over the range of possible scores (i.e., some with scores near the bottom and some with scores near the top), then the method is not working well. What can cause the borderline-group method to work poorly? There are two major causes:

1. The borderline-group may include many students who were put in the group, not because their achievement was actually borderline but because their achievement was difficult for the teachers to judge. (These might be students who have trouble expressing themselves or who are uncooperative.)
2. The teachers may be basing their judgements on something other than what the test measures.

If the spread of scores of the borderline-group is too large, then speak to each teacher individually, making sure that the directions for judging were followed. It is a good idea to find out the names of students judged "borderline" who received outstandingly high or low test scores and ask the teachers to check their classifications on those students. Try not to tell the teachers why you are asking about particular students to avoid the circularity of having the re-judgement based on the test score.

The main advantage of the borderline-group method is that the calculations it requires are very simple. Its main disadvantage is that it uses only a small proportion of all the students taking the test.

If the above procedure is to be used in the process of making important decisions about individual students, it is highly desirable to include as many borderline students as possible (up to 100) to calculate the standard. This can be done either by involving other schools in the district or by accumulating records of borderline scores over a period of time.

5.2 ERRORS OF CLASSIFICATION

A student's score on a test is not a perfect indication of the student's level of mastery. If it was, then the many important decisions involving student progress would be easy to make. The questions on the test are only a small sample of the many quesions that could have been prepared to measure the objectives. A student takes a test at a particular time, on a particular day, under a certain set of conditions. If another test measuring the same objectives were administered on a different day and under different conditions, the student's score would likely be different. The effects of these factors will often be large enough so that some students likely will be misclassified.

You can minimize these errors by ensuring that the test adequately covers the objectives of your course and by ensuring that a maximum number of test items are used to measure each objective.

Perhaps most important, you can minimize errors associated with any single test by ensuring that whenever important decisions are to be made about an individual, results of all tests are combined with your day-to-day observations of the student and his work.

BIBLIOGRAPHY

Bloom, B. S. and others. **Handbook on Formative and Summative Evaluation of Student Learning.** McGraw-Hill, 1971.

Ebel, R. L. **Essentials of Educational Measurement.** Prentice-Hall, 1972.

Gronlund, N. E. **Measurement and Evaluation in Teaching,** 3rd ed. Collier Macmillan Canada, 1976.

Gronlund, N. E. **Stating Objectives for Classroom Instruction,** 2nd ed. Collier Macmillan Canada, 1978.

Mehrens, W. A. and Lehman, I. J. **Measurement and Evaluation in Education and Psychology,** 2nd ed. Holt, Rinehart and Winston, 1975.

Popham, W. J. **Criterion-Referenced Measurement.** Prentice-Hall, 1978.

Thorndike, R. L. and Hagen, E. P. **Measurement and Evaluation in Psychology and Education,** 4th ed. John Wiley and Sons, 1977.

Zierky, M. J. and Livingston, S. A. **A Manual for Setting Standards on the Basic Skills Assessment Tests,** Education Testing Service, 10pp.

A GLOSSARY OF MEASUREMENT TERMS[1]

The terms defined are the more common or basic ones such as occur in test manuals and educational journals. In the definitions, certain technicalities and niceties of usage have been sacrificed for the sake of brevity and, it is hoped, clarity.

academic aptitude The combination of native and acquired abilities that are needed for school learning; likelihood of success in mastering academic work, as estimated from measures of the necessary abilities. (Also called scholastic aptitude, school learning ability, academic potential).

achievement test A test that measures the extent to which a person has "achieved" something, acquired certain information, or mastered certain skills — usually as a result of planned instruction or training.

aptitude A combination of abilities and other characteristics, whether native or acquired, that are indicative of an individual's ability to learn or to develop proficiency in some particular area if appropriate education or training is provided. Aptitude tests include those of general academic ability (commonly classed mental ability or intelligence tests); those of special abilities, such as verbal, numerical, mechanical, or musical; tests assessing "readiness" for learning; and prognostic tests, which measure both ability and previous learning, and are used to predict future performance — usually in a specific field, such as foreign language, shorthand, or nursing.

Some would define "aptitude" in a more comprehensive sense. Thus, "musical aptitude" would refer to the combination not only of physical and mental characteristics but also of motivational factors, interest, and conceivably other characteristics, which are conducive to acquiring proficiency in the musical field.

arithmetic mean A kind of average usually referred to as the mean. It is obtained by dividing the sum of a set of scores by their number.

average A general term applied to the various measures of central tendency. The three most widely used averages are the arithmetic mean (mean), the median, and the mode. When the term "average" is used without designation as to type, the most likely assumption is that it is the arithmetic mean.

diagnostic test A test used to "diagnose" or analyze; that is, to locate an individual's specific areas of weakness or strength, to determine the nature of his weakness or deficiencies, and, wherever possible, to suggest their cause. Such a test yields measures of the components or subparts of some larger body of information or skill. Diagnostic achievement tests are most commonly prepared for the skill subjects.

difficulty value An index which indicates the percent of some specified group, such as students of a given age or grade, who answer a test item correctly.

discriminating power The ability of a test item to differentiate between persons possessing much or little of some trait.

discrimination index An index which indicates the power of a test item to discriminate between higher and lower scoring individuals.

distractor Any incorrect choice (option) in a test item.

distribution (frequency distribution) A tabulation of the scores (or other attributes) of a group of individuals to show the number (frequency) of each score, or of those within the range of each interval.

error of measurement See **standard error of measurement.**

---

[1] Reproduced in part from **A Glossary of Measurement Terms** (Test Service Notebook, No. 13). Distributed by The Psychological Corporation.

f A symbol denoting the frequency of a given score or of the scores within an interval grouping.

formative evaluation Formative evaluation in the classroom is a broad term to encompass all the various evaluative procedures (both formal and informal) conducted periodically during a unit or course for the purpose of identifying areas of students' performance in need of further effort and attention. As well, teachers often use this information to evaluate the effectiveness of instructional procedures, sequencing, illustrative materials, exercises, etc. for purposes of revision and improvement. Students' results on these quizzes and exercises are generally not intended as a method of arriving at a course grade. See also **summative evaluation.**

frequency distribution See **distribution.**

group test A test that may be administered to a number of individuals at the same time by one examiner.

individual test A test that can be administered to only one person at a time, because of the nature of the test and/or the maturity level of the examinees.

internal consistency Degree of relationship among the items of a test; consistency in content sampling.

item A single question or exercise in a test.

item analysis The process of evaluating single test items in respect to certain characteristics. It usually involves determining the difficulty value and the discriminating power of the item, and often its correlation with some external criterion.

Kuder-Richardson formula(s) Formulas for estimating the reliability of a test that are based on inter-item consistency and require only a single administration of the test. The one most used, formula 20, requires information based on the number of items in the test, the standard deviation of the total score, and the proportion of examinees passing each item. The Kuder-Richardson formulas are not appropriate for use with speeded tests.

mastery test A test designed to determine whether a pupil has mastered a given unit of instruction or a single knowledge or skill; a test giving information on what a pupil knows, rather than on how his performance relates to that of some norm-referenced group. Such tests are used in computer-assisted instruction, where their results are referred to as content — or criterion-referenced information.

mean (M) See **arithmetic mean.**

median (Md) The middle score in a distribution or set of ranked scores; the point (score) that divides the group into two equal parts; the 50th percentile. Half of the scores are below the median and half above it, except when the median itself is one of the obtained scores.

multiple-choice item A test item in which the examinee's task is to choose the correct or best answer from several given answers or options.

n The symbol commonly used to represent the number of cases in a group.

normal distribution A distribution of scores or measures that in graphic form has a distinctive bell-shaped appearance. In such a normal distribution, scores or measures are distributed symmetrically about the mean, with as many cases up to various distances above the mean as down to equal distances below it. Cases are concentrated near the mean and decrease in frequency, according to a precise mathematical equation, the farther one departs from the mean. Mean and median are identical. The assumption that mental and psychological characteristics are distributed normally has been very useful in test development work.

norms Statistics that supply a frame of reference by which meaning may be given to obtained test scores. Norms are based upon the actual performance of pupils of various grades or ages in the standardization group for the test. Since they represent average or typical performance, they should not be regarded as standards or as universally desirable levels of attainment. The most common types of norms are deviation IQ, percentile rank, grade equivalent, and stanine. Reference groups are usually those of specified age or grade.

objective test A test made up of items for which correct responses may be set up in advance; scores are unaffected by the opinion or judgement of the scorer. Objective keys provide for scoring by clerks or by machine. Such a test is contrasted with a "subjective" test, such as the usual essay examination, to which different persons may assign different scores, ratings, or grades.

percentile (P) A point (score) in a distribution at or below the percent of cases indicated by the percentile. Thus a score coinciding with the 35th percentile (P_{35}) is regarded as equalling or surpassing 35 percent of the persons in the group. It also means that 65 percent of the performances exceed this score. "Percentile" has nothing to do with the percent of correct answers an examinee makes on a test.

percentile band An interpretation of a test score which takes account of the measurement error that is involved. The range of such bands, most useful in portraying significant differences in battery profiles, is usually from one standard error of measurement below the obtained score to one standard error of measurement above it.

percentile rank (PR) The expression of an obtained test score in terms of its position within a group of 100 scores; the percentile rank of a score is the percent of scores equal to or lower than the given score in its own or in some external reference group.

power test A test intended to measure level of performance unaffected by speed of response; hence one in which there is either no time limit or a very generous one. Items are usually arranged in order of increasing difficulty.

practice effect The influence of previous experience with a test on a later administration of the same or a similar test; usually an increased familiarity with the directions, kinds of questions, etc. Practice effect is greatest when the interval between testings is short, when the content of the two tests is identical or very similar, and when the initial test-taking represents a relatively novel experience for the subjects.

predictive validity See **validity** (2).

profile A graphic representation of the results on several tests, for either an individual or a group, when the results have been expressed in some uniform or comparable terms (standard scores, percentile ranks, grade equivalents, etc.). The profile method of presentation permits identification or areas of strength or weakness.

range For some specified group, the difference between the highest and the lowest obtained score on a test; thus a very rough measure of spread or variability, since it is based upon only two extreme scores. Range is also used in reference to the possible spread of measurement a test provides, which in most instances is the number of items in the test.

raw score The first quantitative result obtained in a scoring test. Usually the number of right answers, number right minus some fraction of number wrong, time required for performance, number of errors, or similar direct, unconverted, uninterpreted measure.

readiness test A test that measures the extent to which an individual has achieved a degree of maturity or acquired certain skills or information needed for successfully undertaking some new learning activity. Thus a readiness test indicates whether a child has reached a developmental stage where he may profitably begin formal reading instruction. Readiness tests are classified as prognostic tests.

recall item A type of item that requires the examinee to supply the correct answer from his own memory or recollection, as contrasted with a recognition item, in which he need only identify the correct answer.

Columbus discovered America in the year _____ is a recall (or completion) item.
See **recognition item.**

recognition item An item which requires the examinee to recognize or select the correct answer from among two or more given answers (options).

Columbus discovered America in

(a) 1425 (b) 1492 (c) 1520 (d) 1546

is a recognition item.

reliability The extent to which a test is consistent in measuring whatever it does measure; dependability, stability, trustworthiness, relative freedom from errors of measurement. Reliability is usually expressed by some form of reliability coefficient or by the standard error of measurement derived from it.

reliability coefficient The coefficient of correlation between two forms of a test, between scores on two administrations of the same test, or between halves of a test, properly corrected. The three measure somewhat different aspects of reliability, but all are properly spoken of as reliability coefficients.

skewed distribution A distribution that departs from symmetry or balance around the mean, i.e., from normality. Scores pile up at one end and trail off at the other.

standard deviation (S.D.) A measure of the variability or dispersion of a distribution of scores. The more the scores cluster around the mean, the smaller the standard deviation. For a normal distribution, approximately two thirds (68.3 percent) of the scores are within the range from one S.D. below the mean to one S.D. above the mean. Computation of the S.D. is based upon the square of the deviation of each score from the mean. The S.D. is sometimes called "sigma" and is represented by the symbol σ.

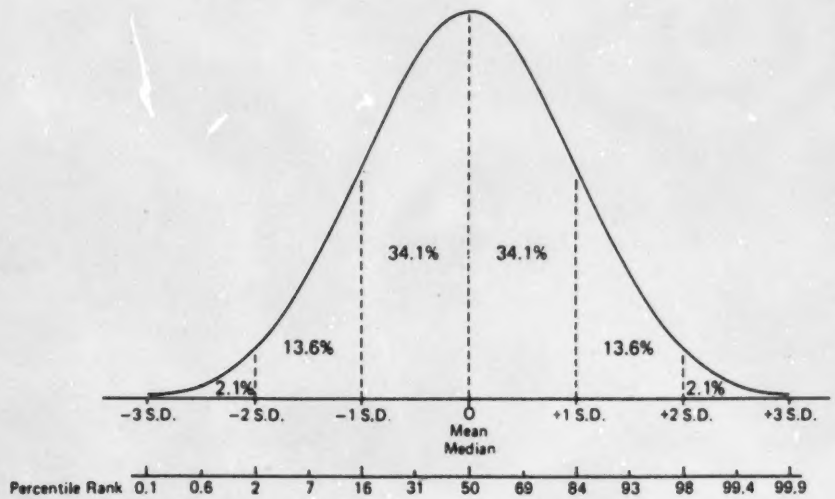


Figure 1. Normal curve, showing relations among standard deviation from mean, area (percentage of cases) between these points and percentile rank.

standard error (S.E.) A statistic providing an estimate of the possible magnitude of "error" present in some obtained measure, whether (1) an individual score or (2) some group measure, such as a mean or a correlation coefficient.

(1) **standard error of measurement (S.E.M.):** As applied to a single obtained score, the amount by which the score may differ from the hypothetical true score due to errors of measurement. The larger the S.E.M., the less reliable the score. The S.E.M. is an amount such that in about two-thirds of the cases the obtained score would not differ by more than one S.E.M. from the true score. (Theoretically, then, it can be said that the chances are 2:1 that the actual score is within a band extending from true score minus 1 S.E.M. to true score plus 1 S.E.M.; but since the true score can never be known, actual practice must reverse the true-obtained relation for an interpretation.) Other probabilities are noted under (2) below. See **true score.**

(2) **standard error:** When applied to group average, standard deviations, correlation coefficients, etc., the S.E. provides an estimate of the "error" which may be involved. The group's size and the S.D. are the factors on which these standard errors are based. The same probability interpretation as for S.E.M. is made for the S.E.'s of group measures, i.e., 2:1 (2 out of 3) for the 1 S.E. range, 19:1 (95 out of 100) for a 2 S.E. range, 99:1 (99 out of 100) for a 2.6 S.E. range.

standard score A general term referring to any of a variety of "transformed" scores, in terms of which raw scores may be expressed for reasons of convenience, comparability, ease of interpretation, etc. The simplest type of standard score, known as a z-score, is an expression of the deviation of a score from the mean score of the group in relation to the standard deviation of the scores of the group. Thus:

$$\text{standard score (Z)} = \frac{\text{raw score (X) - mean (M)}}{\text{standard deviation (S.D.)}}$$

Standard scores are useful in expressing the raw scores of two forms of a test in comparable terms in instances where tryouts have shown that the two forms are not identical in difficulty; also, successive levels of a test may be linked to form a continuous standard-score scale, making across-battery comparisons possible.

standardized test (standard test) A test designed to provide a systematic sample of individual performance, administered according to prescribed directions, scored in conformance with definite rules, and interpreted in reference to certain normative information. Some would further restrict the usage of the term "standardized" to those tests for which the items have been chosen on the basis of experimental evaluation, and for which data on reliability and validity are provided. Others would add "commercially published" and/or for "general use".

stanine One of the steps in a nine-point scale of standard scores. The stanine (short for standard-nine) scale has values from 1 to 9, with a mean of 5 and a standard deviation of 2. Each stanine (except 1 and 9) is ½ S.D. in width, with the middle (average) stanine of 5 extending from ¼ S.D. below to ¼ S.D. above the mean. (See Figure 2.)

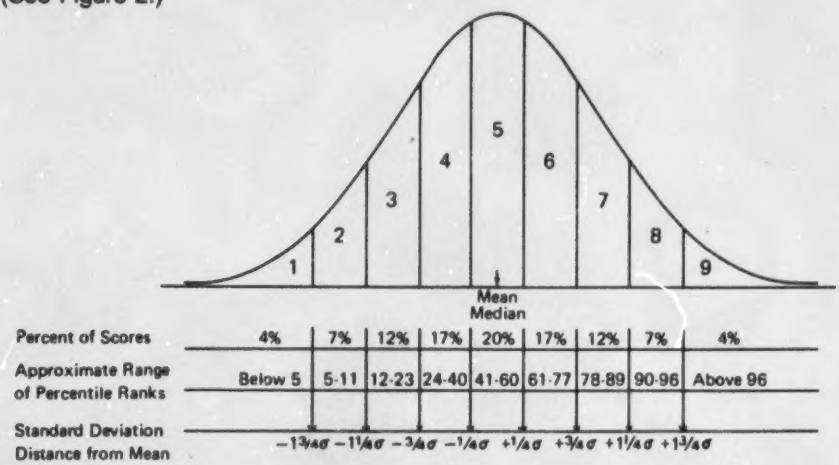


Figure 2. Stanines and the normal curve. Each stanine (except 1 and 9) is one half S.D. in width.

summative evaluation Summative evaluation in the classroom generally is used to refer to evaluative procedures (tests, examinations, reports, projects) conducted customarily at the end of major units for the purpose of assessing a student's performance in relation to others and/or to a predetermined criterion level. These results normally form a substantial basis for the student's course grade. An additional purpose of summative evaluation is to provide the teacher with information concerning the relative effectiveness of the preceding unit in meeting the designated instructional objectives. See also **formative evaluation.**

taxonomy An embodiment of the principles of classification; a survey, usually in outline form, such as a presentation of the objectives of education.

true score A score entirely free of error; hence, a hypothetical value that can never be obtained by testing, which always involves some measurement error. A "true" score may be thought of as the average score from an infinite number of measurements from the same or exactly equivalent tests, assuming no practice effect or change in the examinee during the testings. The standard deviation of this infinite number of "samplings" is known as the standard error of measurement.

validity The extent to which a test does the job for which it is used. This definition is more satisfactory than the traditional "extent to which a test measures what it is supposed to measure," since the validity of a test is always specific to the purposes for which the test is used. The term validity, then, has different connotations for various types of tests and, thus, a different kind of validity evidence is appropriate for each.

(1) **content, curricular validity** For achievement tests, validity is the extent to which the content of the test represents a balanced and adequate sampling of the outcomes (knowledge, skills, etc.) of the course or instructional program it is intended to cover. It is best evidenced by a comparison of the test content with courses of study, instructional materials, and statements of educational goals; and often by analysis of the processes required in making correct responses to the items. Face validity, referring to an observation of what a test appears to measure, is a non-technical type of evidence; apparent relevancy is, however, quite desirable.

(2) **criterion-related validity.** The extent to which scores on the test are in agreement with (concurrent validity) or predict (predictive validity) some given criterion measure. Predictive validity refers to the accuracy with which an aptitude, prognostic, or readiness test indicates future learning success in some area, as evidenced by correlations between scores on the test and future criterion measures of such success (e.g., the relation of score on an academic aptitude test administered in high school to grade point average over four years of college). In concurrent validity, no significant time interval elapses between administration of the test to one generally accepted as or known to be valid, or by the correlation between scores on a test and criteria measures which are valid but are less objective and more time-consuming to obtain than a test score would be.

(3) **construct validity.** The extent to which a test measures some relatively abstract psychological trait or construct; applicable in evaluation the validity of tests that have been constructed on the basis of an analysis (often factor analysis) of the nature of the trait and its manifestations. Tests of personality, verbal ability, mechanical aptitude, critical thinking, etc., are validated in terms of their construct and the relation of their scores to pertinent external data.

variability. The spread or dispersion of test scores, best indicated by their standard deviation.